University of Colorado, Boulder
CU Scholar

School of Education Graduate Theses & Dissertations

School of Education

Spring 1-1-2011

A Validity Study of Interim Assessments in an Urban School District

Elena Kitaoka Diaz-Bilello
University of Colorado at Boulder, Elena.DiazB@Colorado.EDU

Follow this and additional works at: http://scholar.colorado.edu/educ_gradetds

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

Diaz-Bilello, Elena Kitaoka, "A Validity Study of Interim Assessments in an Urban School District" (2011). *School of Education Graduate Theses & Dissertations.* Paper 10.

This Dissertation is brought to you for free and open access by School of Education at CU Scholar. It has been accepted for inclusion in School of Education Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.

A VALIDITY STUDY OF INTERIM ASSESSMENTS IN AN URBAN SCHOOL DISTRICT

by

Elena K. Diaz-Bilello

B.A., Lewis & Clark College, 1993

M.P.A., Columbia University, 2000

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the
Doctor of Philosophy

Research and Evaluation Methodology Program
in the Graduate School of Education
University of Colorado at Boulder
April 2011

This thesis entitled:
A Validity Study of Interim Assessments in an Urban School District
written by Elena Diaz-Bilello
has been approved for the School of Education
at the University of Colorado, Boulder

____
Associate Professor Derek C. Briggs, Committee Chair

____
Assistant Professor Edward W. Wiley, Committee Chair

____
Dean Lorrie A. Shepard, Committee Member

____
Dr. Scott F. Marion, Committee Member

____
Dean Paul Teske, Committee Member

Date_____________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline

HRC protocol # 1008.7

Abstract

Diaz-Bilello, Elena (Ph.D., Education, Research and Evaluation Methodology Program)

A Validity Study of Interim Assessments in an Urban School District

Thesis directed by Associate Professor Derek C. Briggs and Assistant Professor Edward W. Wiley

Despite the large investment and rapid deployment of interim assessments in school districts across the nation, the variability in standards used to develop these tests, and the expectation by users that these assessments provide valid data for evaluative, predictive, and instructional uses, few studies have been conducted to examine whether specific uses of the test can be supported or justified (Perie, Marion and Gong, 2009; Shepard, 2009, 2007; Herman & Baker, 2005).

This dissertation uses Kane's (2006) argument-based approach to evaluate whether instructional, predictive, and evaluative uses of interim assessments in the Denver Public Schools are supported. The evaluation consists of using both quantitative and qualitative approaches to test out the inferences and assumptions specified in the interpretive argument.

While there is a clear trend in school districts across the country to use interim assessments, the key findings from this study reveal that these assessments may not always be providing valid information to drive aspects of accountability and reforms such as evaluating teacher effectiveness, school performance, and improving instruction.

Acknowledgements

This dissertation could not reach completion without the invaluable feedback, support, sense of humor and compassion from numerous people and canine friends. I am honored to acknowledge them here.

I am deeply indebted to my dissertation committee members: Dr. Derek Briggs, Dr. Edward Wiley, Dean Lorrie Shepard, Dr. Scott Marion, and Dean Paul Teske. Special thanks to my dissertation chairs, Dr. Briggs and Dr. Wiley, for their generous time and countless reviews of drafts. In addition, I am grateful to the following members of the School of Education for their invaluable help to me during this long dissertation journey: Jon Weeks, Marie Huchton, Dr. Margaret Eisenhart, Dr. Daniel Liston, Dr. Kenneth Howe, and Dr. Guillermo Solano-Flores.

I hold in deep gratitude, the unfailing support from my family, friends, and colleagues. Eternal thanks in particular to: Daniel Bilello, Cheyenne, Jun Diaz, Temma Shoaf, Reuven Diaz, Jeni Glasgow, Steve Lehman, Henry Roman, Shirley Scott, Zu Zu, Einar, and Floyd. In addition, I am grateful for the support and the companionship of the many friends with whom I have skied and hiked the great mountains of Colorado.

Lastly, I am tremendously grateful to the staff and teachers of the Denver Public Schools for their time and willingness to participate in this study, and to the Piton Foundation for providing partial funding support.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 - Introduction

Largely in response to increased accountability requirements brought about by the inception of the federally mandated No Child Left Behind Act (NCLB) in 2001, many school districts in the United States have invested in interim assessments as tools for improving and tracking student achievement (Bulkley, Christman, Goertz and Lawrence, 2010; Perie, Marion and Gong, 2009; Cech, 2008; Herman, Yamashiro and Lefkowitz, 2008; Vendlinski, Nagashima, and Herman, 2007; Shepard, 2007; Herman & Baker, 2005; Olson, 2005). According to Olson (2005), one out of seven superintendents surveyed nation-wide indicated that they either used, or were planning to institute, interim assessments in their school district. These testing systems consist of formal assessments, often referred to as “interim”, “periodic”, “formative”, or “benchmarks” in different school districts, that are administered to students at least two times a year to provide data on individual student performance. The standardized format of these tests allow the data from these assessments to be rolled up and examined at different levels of aggregation (e.g., by classrooms, by school, or by district).

Interim assessments range in form from comprehensive software and reporting packages, such as the Northwest Evaluation Association’s *Measures of Academic Progress* system, to tests developed by individual school districts from item banks created by companies such as CTB-McGraw Hill, Pearson, and the Riverside Testing Company (Olson, 2005; Cech, 2008). Regardless of whether interim assessments are developed in-house by a school district or purchased directly from a vendor, these assessments require substantial time and resources to acquire, develop, administer, and report each year. District policymakers and assessment staff justify investments in interim assessment systems on the basis that the tests provide them with important data to inform decision-making in, or across, classrooms and schools, and that these products can help improve student achievement (Perie et al., 2009; Cech, 2008; Herman, et al., 2008; Shepard, 2007; Herman & Baker, 2005).

The relationship between interim assessments and improved student achievement is commonly emphasized by test vendors in the marketing of their products to school districts and by school districts justifying the use of interim assessments to stakeholders. In promoting the benefits of interim assessment systems, both test vendors and school districts typically draw on terms used in research on classroom teaching and learning to strengthen the connection between their products and improved student learning and achievement in the classrooms. For example, one vendor, Renaissance Learning, states, “with education budgets tighter than ever, you need to make every dollar count. That’s why more teachers trust Accelerated Reader over any other reading program, to provide frequent progress monitoring and produce the greatest reading improvement for the least investment”. Renaissance Learning claims that the term, Zone of Proximal Development (ZPD), introduced by Vygotsky (1978) to describe a child’s cognitive development relative to tasks performed independently or through social interactions with others, can be measured by their tests. According to Renaissance Learning, the quantified ZPD level represents "the level of difficulty that is neither too hard nor too easy, and is the level at which optimal learning takes place". By claiming that a child’s ZPD can be measured and quantified through their assessment system, Renaissance Learning maintains that students who are encouraged to read within their ZPD level will subsequently improve their vocabulary and comprehension skills.

In another example, a competitor to Renaissance Learning, The Princeton Review (TPR), draws extensively on the research of Black and William (1998) to provide evidence for consumers on the benefits of formative assessments and their formative assessment products on classroom learning. After outlining the benefits of formative assessments by citing key findings from Black and William’s paper, the company states that, “In this climate of accountability, educators need a tool that:

- Helps students reach proficiency and then continue to mastery.
- Closes the achievement gap between subgroups of students.
- Uses high-quality data to track student progress and chart their paths to success.”

According to TPR, when data from their formative assessments are used as a student monitoring and evaluation tool, their assessments can help users improve student learning in the classroom and meet all three of the above-stated goals. Although the extent to which TPR products appear to share the same qualities as the classroom assessments described by Black and William is not discussed, the company's use of the term "formative" is intended to let consumers know that the product can result in the same practices and learning gains reported by Black and William.

School districts, like testing companies, also draw on the research base from classroom teaching and learning to justify their investment in interim test systems. For example, in 2008, Miami Dade County Public Schools District released a research brief outlining the benefits of investing in and using an interim assessment system. In the research brief, the author acknowledges that although, "researchers have noted that many commercially developed interim assessments claim to be formative but are only mini-standardized tests intended to predict how well students will perform on end-of-year state tests", the author at the same time draws on the formative assessment literature to support the use of a "formative interim assessment" in Miami Dade County. In Miami-Dade County, the multiple-choice interim assessment system was developed collaboratively between district staff and a testing company, and the assessments were designed to both follow the "district's pacing guides…and comply with FCAT[1] passage and item specifications" (2008, pg 6).

Although entities such as Renaissance Learning, TPR and school districts using interim assessments uniformly claim that these products improve classroom instruction and learning, interim assessments, unlike their high-stakes end of year or summative test counterpart, represent a range of products developed using variable standards of processes and procedures. As documented by the New Mexico Public Education Department (2006) in their review of interim assessment test products, the extent to which these products meet technical criteria, state standards and expectations, and the needs of special populations can vary greatly. Villano (2006) also notes that some interim assessment products are populated with items that aligned with a given state's standards, whereas other products are populated with items that have not been updated for many years. Further, the standards and process for determining what constitutes "proficiency" or student mastery over content being measured varies between interim

[1] FCAT or the Florida Comprehensive Assessment Test is the high-stakes end of year summative assessment administered to all students in Florida.

assessments. For example, vendors such as the Northwest Evaluation Association (NWEA) and Scantron use computer adaptive tests consisting of items with varying difficulty to determine the proficiency of a student based on the student's responses, whereas, in other cases, test vendors such as TPR, who provide items to school districts allow their clients to set the proficiency standards for each test. Although few educators would argue against evaluating and assessing student learning, the variability of standards and processes used to develop interim assessments may pose significant problems to consumers depending on how these tests are being used. Further, any assessment used to evaluate students or teachers with high consequences would warrant more scrutiny to ensure that those uses can be justified.

Interim assessment uses in schools and school districts typically encompass three categories: evaluative, predictive and instructional (Perie, et al., 2009). Schools, districts, and states employing interim assessments for evaluative purposes use these data to evaluate and monitor the effectiveness of programs or reforms. The predictive use of the interim assessments refers to the practice of using interim assessments to predict the performance of students on the end of year high stakes assessment and to structure interventions that may boost the achievement of students identified as low performers prior to the high-stakes testing window (Perie et al., 2009, Shepard, 2009). The final category, instructional, refers to the use of interim assessments to either drive instructional decisions or change instructional practices within the classroom. Depending on the state or district, interim assessment uses may not be restricted to just one category and may encompass one, two, or all three of the categories identified by Perie, et al.

Despite the large investment and rapid deployment of interim assessments in school districts across the nation, the variability of standards used to develop these tests, and the expectation by users that these assessments provide valid data for evaluative, predictive, and instructional uses, few studies have been conducted to examine whether specific uses of the test can be supported or justified (Perie et al., 2009; Herman et al., 2008; Shepard, 2007; Herman & Baker, 2005). Shepard (2009) states, "interim assessments could sometimes be a good thing, but they are brand new and wholly unexamined. Therefore, some amount of skepticism and search for evidence is warranted" (pg. 35). Although measurement experts beginning with Cronbach (1988) and later with Shepard (2007, 1993), Kane (2006, 1992) and Wilson (2004) have long advocated for the use of an evaluative framework to validate tests, limited analysis has been conducted at the district or state level to establish policy or procedures to evaluate and validate the uses of interim tests and the claims made about what these tests can achieve.

Purpose Statement

The purpose of this dissertation is to evaluate whether instructional, predictive, and evaluative uses of interim assessment were supported by an interim assessment program customized for a large urban school district by a testing company during the 2006-07 and 2007-08 school years. In developing these tests for the school district, item writers hired by the testing company wrote the items and district personnel set the cut-points for determining student proficiency levels. All tests were administered to students district-wide before the items were piloted, and used by this school district to support decisions directly impacting both teachers and students. Although this dissertation focuses on one interim assessment program used in a large urban school district, the process and procedures used to develop the interim tests in this school district are not unique to this particular school district. That is, at the national level, other school districts customize their interim assessments through the same testing company, and many other districts undergo a similar process of creating and rapidly administering interim assessments through other testing vendors offering customized services. In the context of this national trend, insights from this district example can shed light on assumptions and findings that may be relevant to many other districts across the country.

To evaluate interim assessments relative to specific uses, this study adopts the interpretive argument framework articulated by Kane (2006). This framework first requires identifying the areas or the qualities of a test that one would argue must meet certain criteria in order for the test system to plausibly support specified uses. After identifying the inferences and assumptions that support the specified uses of the test, the study evaluates whether the selected arguments supporting the inferences made about the test system can withstand scrutiny. The final stage of the study requires evaluating the entire interpretive argument to determine whether the evidence gathered sufficiently supports the specified uses of the test.

Background of Case Study District

The district examined in this study, DPS, is the largest urban school district in the state of Colorado with approximately 73,000 students enrolled annually. As is typical of many urban school districts, a large concentration of high-poverty schools (schools with over 70% of students participating in the free and reduced lunch program) deemed to be "low" performing schools by the state accountability rating system reside within this district. DPS also serves a larger population of students of color (56.3% Latino, 18% African American, 1% American Indian, and 3% Asian in the 2007-08 school year) relative to other districts in the state. In Colorado, no other school district shares the same scale of poverty and low performing schools as DPS. Due to the district's large number of underperforming schools facing an array of state, federal and district accountability sanctions, combined with the presence of other challenges such as high dropout and low graduation rates for students of color, DPS has faced tremendous public scrutiny and pressure to improve the educational system and student achievement. In response to public pressure to improve, multiple educational reforms and programs have been put in place in DPS over the years specifically for the purpose of raising student achievement.

Within DPS, the interim assessment system was built in response to a mandate issued by the Superintendent and the Chief Academic Officer (CAO) to better monitor student achievement and to improve instructional quality across the district. Both the Superintendent and the CAO of DPS hoped that such assessments would provide more data to monitor student achievement and growth over the school year as a means of predicting student performance on high-stakes CSAP tests (DPS assessment staff, personal communication, 2007). For many schools and districts in Colorado, predicting student performance on the CSAP is important to identifying which schools may face state-imposed sanctions such as conversion from public to a charter school entity or federally-imposed sanctions from NCLB legislation. Federally imposed sanctions from NCLB range from having to determine how many students will take advantage of “choosing out” of a low-performing school each year to potentially dismissing and re-hiring the instructional leader or all staff at a given school. In the 2007-08 school year, out of 184 schools in the district, 31 schools faced various “restructuring” measures (with 8 schools identified for school closure and 5 schools re-opening with entirely new staff and curricula in place); 12 schools faced “corrective action”; and, 25 schools were placed on “school action” plans. Since more than half of all schools in DPS are either rated as “low” performing schools by the state each year or do not make “adequate yearly progress” under NCLB, policymakers are especially interested in monitoring the performance of these schools in anticipation that some schools may face sanctions and require district intervention.

The DPS interim assessments currently consist of “benchmark” tests administered to grades 3 through 8 which target the three content domains (reading, writing, and math) measured by the high-stakes CSAP tests and “course assessments” which cover content taught in core subject areas for grades 9 and 10. In contrast to the CSAP which is administered only once in during the month of March, all interim assessments in the two years of data (2006-07 and 2007-08 school years) reviewed for this dissertation were administered during three separate time points in the school year: the beginning of the fall semester (September); at the end of the fall semester (December); and, at the end of the school year (May).

In the 2006-07 and 2007-08 school years, DPS developed the interim assessment system in partnership with The Princeton Review (TPR) and Edusoft. That is, item panels from DPS reviewed all of the items developed by TPR for each assessment and made modifications to the reading passages and items as needed. This process of having district stakeholders provide direct input on the items used for each test in not unique to this district, and parallels the process instituted in other districts such as in Philadelphia, by the same test vendor. At DPS, the decision to work with TPR and Edusoft was motivated largely by the flexibility provided by the vendors to allow district staff to modify items as they wished and to request items that directly addressed the state content standards targeted by each assessment. The role of Edusoft was confined to providing the reporting system that enabled all teachers to access student data immediately after the tests were scanned by each school. TPR was directly involved with the development of all tests and provided DPS with test items that were then either modified or re-written by district staff members. Although DPS continues to use some of the items purchased from TPR for the interim tests administered to grade 3-8 students, the district developed an entirely separate set of items in 2007-08 to be used as course assessments in the high schools. These course assessments differ from the benchmarks in that these assessments tested other areas, such as Biology and Geology, in addition to subjects related to reading, writing, and math. In addition, since not all items from TPR were re-used in the current set of interim tests, district staff developed many new items for the grades 3-8 interim tests in the 2008-09 school year.

Based on the attributes and context described for this one large urban school district, the findings from this case study would ideally provide useful insights to other school districts facing similar challenges across the nation. Since many urban school districts face similar accountability pressures as DPS, these districts would most likely be enacting similar practices around interim assessment use and may benefit from findings shared in this study.

Research Questions

Based on the common practice found across school districts to use data from interim assessment to drive decisions that have both high and low consequences on users and test-takers, the larger research question in this study asks: to what extent can stakeholders draw on data from interim assessments to inform and drive predictive, instructional, and evaluative uses of the test? To explore this question, the study focuses on four different uses of the test that span the three "use" categories identified by Perie et al. (2009) in one large urban school district. Evidence in this study was gathered to learn the extent to which the interim tests can:

1. Predict the performance on the high-stakes summative state assessments for students before the high-stakes testing period;

2. Identify low-performing students for mandatory summer remediation;

3. Provide adequately accurate measures of student growth for the purpose of evaluating teacher performance; and,

4. Be used as an important tool by teachers to improve instruction by meeting the learning needs of individual students.

The first use reflects a common instructional use of many formal and informal assessments in schools. That is, students who exhibit poor or low performance on an assessment would be identified by teachers for remediation either within class or during after-school tutoring. The first use reflects a commonly practiced use of interim tests as an early warning system to predict student performance on the high-stakes state assessment. In the case study district, each school site implemented and structured their own set of interventions to improve the achievement of low performing students on the interim assessments. Although some students may not need remediation, the consequences may be minimal and some students may stand to benefit from receiving extra instructional assistance. District staff and a few of the interviewed teachers indicated that many schools used the data to identify low performing students for remediation or tutoring and some schools organized students into ability groups to receive different instructional interventions.

The second use represents a one-time use made district-wide during the first year of implementing the interim assessments in 2006-07 to identify low achieving grade 8 students on the interim assessments for mandatory summer remediation, thereby raising the stakes for those grade 8 students taking the reading and math interim tests that year.

The third use presents an evaluative use of the test that will become more common in school districts across the country as more states indicate plans to use interim assessments in their evaluation of teacher effectiveness (Buckley & Marion, 2011). In this particular school district, a teacher can earn a merit pay bonus and a salary increase based on the amount of growth achieved by students between the first and third interim assessments. The last use examined in this study represents an instructional use of the tests as valuable tools for helping teachers improve and modify their instruction to meet individual student needs in the classroom. As noted earlier, both testing companies and school districts commonly claim that these tests serve an instrumental purpose in improving the quality of instruction and student achievement.

The methods of finding evidence to support the validity of the first three uses consist of technical approaches that evaluate the properties of the interim tests against the properties and standards established by the test vendor of the high-stakes Colorado States Assessment Program (CSAP). Due to the absence of absolute external standards for evaluating a given assessment system, this study evaluates the properties of individual items and the reliability of interim assessments in the case study district against the standards employed by the CSAP. Although some interim assessments are created independently of a high-stakes assessment, many test vendors marketing interim assessment products make a concerted effort to demonstrate that these tests are correlated with an end of year summative assessment given in different states. In this case study district, although it may not be reasonable to expect that the substantially shorter interim assessments display the exact same technical qualities as the CSAP, the standards used by the CSAP serves as a logical frame of reference for assessing tests modeled after their high-stakes counterpart. Other educational researchers drawing on the standards used by another established assessment include Vendilinski et al. (2007), who evaluated their newly developed science interim assessments in California using the standards used by the high-stakes state assessment system.

In contrast to the approach used to evaluate the validity of the first three uses, the last use evaluated in this study draws on the input from a group of DPS veteran teachers using the interim assessments and district-wide survey responses from teachers to explore how interim tests are improving instruction in the classrooms. Although these findings do not provide a confirmatory picture of the extent to which these tests can improve classroom instruction, the findings provide insights that are suggestive of the possible benefits and limitations of these assessments within the classroom setting from the perspective of teachers.

Study Overview

This study begins in Chapter 2 with background information on the test development process used by the case study district to create the interim tests and establish the cut-scores to designate the proficiency level of students.

Chapter 3 provides the rationale for using an interpretive argument framework within the larger historical context of test validity and maps out the interpretive argument used in this study to evaluate the specified uses of the interim tests. In this chapter, the interpretive argument presented by Kane (2006) to make trait inferences using an assessment is modified to fit the context of evaluating the interim assessments for three of the four uses. A separate argument is created to address the use of the test for improving instruction. The two distinct arguments presented in this chapter map out the sequence of studies undertaken to evaluate the assumptions supporting the inferences in subsequent chapters.

Chapters 4 and 5 are dedicated to evaluating the set of assumptions supporting the first interpretive argument. The inferences and assumptions in the first argument pertain to the uses of the tests to predict student outcomes on the state test, identify students for summer remediation, and as a measure of teacher effectiveness for merit pay purposes. Chapter 4 evaluates the set of shared inferences and assumptions among the three uses, and Chapter 5 is dedicated to evaluating the additional assumptions under merit pay. The assumptions specific to the merit pay use are evaluated in a separate chapter since the analyses are based entirely on classroom level performance outcomes.

The final set of analyses presented in Chapter 6 evaluate whether teachers are using the assessments to improve their instructional practices. A separate interpretive argument was constructed for this particular use because the claim evaluates the perceptions and beliefs of a group of teachers and moves away from the technical properties evaluated in the previous two chapters. The focal point of each analysis in Chapter 6 was to determine the extent to which teachers were able to make meaningful connections with the interim assessment data to help transform and improve their instructional practices within the classroom setting.

This study culminates with a discussion in Chapter 7, that assesses the evidence gathered from evaluating the interpretive arguments in Chapter 4 through 6. This last chapter ties the relevance of this case study back to the broader issue regarding the current national trend to utilize these types of assessments to drive important decisions such as evaluating teacher performance for merit pay.

CHAPTER 2 - Developing the Interim Test Program

The interim tests evaluated in this study consist of items provided to the Denver Public Schools (DPS) by The Princeton Review (TPR). The study restricts the evaluation to the first two school years (2006-07 and 2007-08) when TPR collaborated with DPS to develop the interim test system. In the 2006-07 school year, all of the items used for each of the grades 3 through 10 tests were customized for the district by TPR. These items were written in order to address the specifications in the blueprints created by district assessment and content specialists. In 2007-08, all test items produced for the grades 3 through 8 tests still came from TPR, but the grades 9 and 10 interim tests were developed internally by district staff. From the 2008-09 school year to the present date, although some of the items used in the grades 3 through 8 interim tests in 2007-08 were re-used, many of the items have been replaced by new items developed by DPS staff.

This chapter provides an overview of the process undertaken by the school district to develop, administer, and score the interim tests over the two year period reviewed. The chapter also concludes with a brief comparison of the process used to develop the DPS interim tests relative to the standards and process used by the vendor who created the state tests that these interim tests were modeled after. As indicated earlier in the first chapter, the process taken by the district to develop these interim tests is not unique to this case study district and largely reflects the experiences of other school districts that chose to or currently work with TPR. Districts typically seek to work with this vendor because district personnel retain autonomy over deciding which items belong in the assessment and are largely responsible for content and determining appropriate uses for the tests.

The Test Development Process

Designing the Blueprints

In order for TPR to customize items for DPS, both entities agreed that the school district would develop the blueprints specifying the number of items required and the content that the items would measure on each test, and that TPR would furnish items meeting the specifications of the blueprint. DPS assessment staff and content experts in the curriculum department developed the interim assessment blueprints with the following considerations factored into the design: that a few constructed-response (CR) items were included on each test to provide teachers with an opportunity to evaluate student work; that the items were representative of grade level standards and expectations covered in the classroom; and, that the tests measured specific standards (power standards) and benchmarks (referred to in the scoring reports as CSAP framework statements) that were also measured by the high-stakes Colorado State Assessment Program (CSAP). The blueprints created addressed the content to be measured for each of the three testing periods during the school year. To ensure that the tests properly measured student knowledge of grade-level math, reading and writing content at each time point, curriculum experts and assessment staff collaborated closely to align the standards and CSAP frameworks reflected in the blueprints with the pacing guide used by teachers for lesson planning purposes.

In addition to aligning both standards and framework statements with the pacing guides, assessment staff limited the number of items used on each test to minimize the amount of time needed to administer and score these tests during the school year. In the 2006-07 year, the blueprints called for 18 items for all reading, math and writing tests. In the 2007-08 school year, assessment staff added more items to the blueprints for the reading and math test, with seven more items added to reading and five more items added to the math tests. Additional items were added to these two tests in 2007-08 based on feedback received from schools that more items were needed to provide better measures of the larger standards and framework statements represented on the tests. Unlike the reading and math tests, writing consisted of 18 items in 06-08 and 17 items in 2007-08. For writing, only one item in the test consisted of a CR item or a writing prompt. For reading and math, three to four items out of the total number of items used for each test consisted of CR items.

Item Selection and Panels

After DPS staff developed the blueprints for the interim test system, TPR item writers developed items as specified by the blueprints. While the items were being customized for the district, assessment and curriculum staff recruited teachers to join the item panels. The panels consisted of curriculum experts working in the central office and teachers identified by central staff members as content experts. Items reviewed by panel members were delivered by TPR in the form of a test booklet and item members were asked to evaluate each item and determine whether the item required modification or needed to be replaced. This review of items included evaluating the scoring rubrics to be used by teachers for rating responses to CR items.

Panels convened three times before the tests were published prior to each test administration. The first set of meetings took place for a full week and entailed a review of all items provided by TPR for grades 3 through 10 in each content area. Following the weeklong review, any modifications made to the items were provided to TPR. The second set of meetings took place over a three-day period and entailed ensuring that all requested changes were addressed by TPR. The last round of reviews consisted of a three day check on all booklets to ensure that panel members approved of the final draft prior to publication. The total time dedicated by item panels to evaluating items and approving the final draft of each test spanned approximately two weeks. However, the time period prior to production varied for each test, particularly if a test required many replacement items. In the case of item replacement, TPR item writers wrote new items addressing the specifications and concerns raised by item panel members. These items, in turn, had to be re-examined and reviewed by item members and additional modifications could be requested to revise the new items[2].

Initially, the contract with TPR required that a representative of the company facilitate the item panel discussions. However, since assessment staff believed at that time, that panel members did not receive adequate guidance from the TPR representative for evaluating each item provided by the

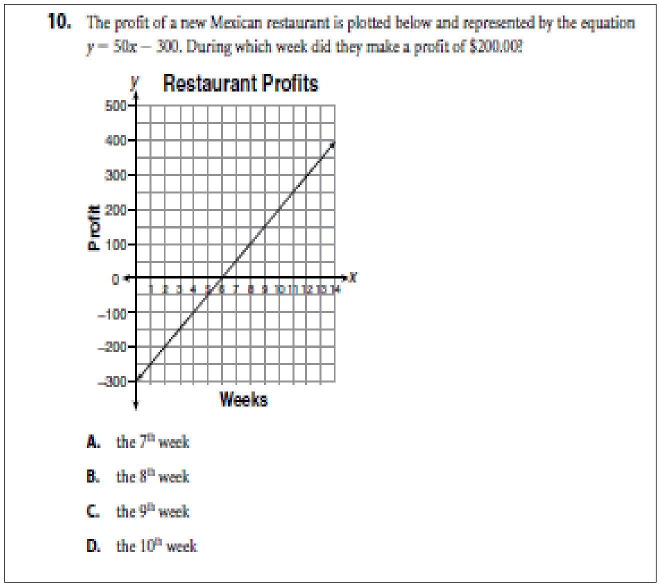
---

[2] In one instance, all items on an elementary math test had to be replaced since panel members agreed that the content and standards reflected middle school expectations for students.

company[3], the assessment staff decided to lead the panel discussions and provide members with criteria for evaluating the quality of the customized items. Criteria for evaluating items were created to help members detect and address language that may be considered to be culturally biased or inappropriate, identify language that may mislead students into selecting or writing an incorrect response, and evaluating the difficulty of an item. For the first two criteria, assessment staff provided examples of items displaying bias and different types of distracters to help panel members conceptualize and become familiar with item characteristics that should be avoided. The third criterion, identifying the difficulty of the item, was evaluated by asking panel members to use descriptors established by the CSAP test vendor of how students at varying proficiency levels typically perform on the test, and to match the proficiency descriptors against the task required by the item. Items were subsequently classified as easy if they could be answered correctly by a "typical" unsatisfactory performing student and very difficult if panel members judged that mostly advanced students could respond correctly. Formal documents recording the panel member's final judgment of each item in 2006-07 and 2007-08 were not compiled, and the information was only used during the panel meetings to ensure that each test consisted of items measuring different types of students.

Although comparisons of item difficulty statistics relative to the judgment made by panel members cannot be made in this study, Figure 1 presents an example of one MC item and one CR item used in a grade 8 math test to identify low performing students for mandatory summer remediation. The first item (item 10) in Figure 1 represents an easy item. For this problem, students would not need to use the equation included in the prompt to answer this item correctly, and would only need to use the graph to find the corresponding week associated with the profit amount of $200.00 presented on the y-axis. Item 10 represents the easiest item on this test with 81 percent of the entire test-taking population, or 3,230 out of 3,999 students responding correctly to this item. The second item, (item 6), asks students to graph an equation.

[3] According to district assessment staff, the TPR representative limited discussions of evaluating the item quality to questions about whether panel members "liked" or "did not like" an item.



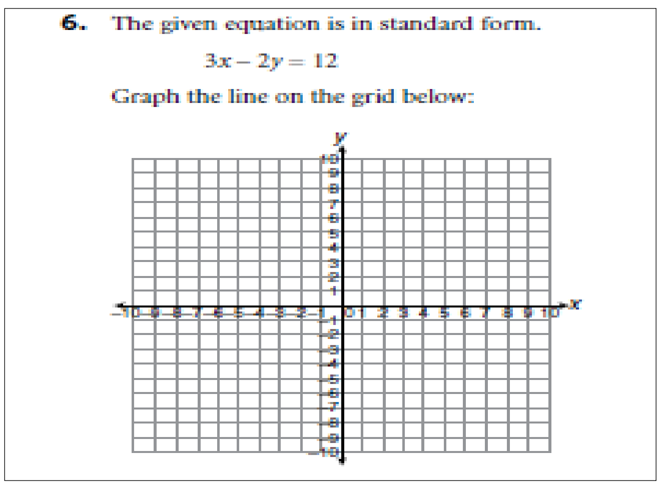


Figure 1. An "easy" and a "difficult" item on the grade 8 math, version 2, 2006-07 test. (Reprinted with permission from the Denver Public Schools)

According to a math specialist in the district, this item represents one of the few "advanced" items used in this test to determine whether a grade 8 student could be waived from taking Algebra I offered in grade 9. For this item, the likelihood of students receiving partial credit or full credit for this item was relatively low. For item 6, teachers were asked to rate the response using the holistic rubric presented below:

Apply 2-point question holistic rubric

2 points –
Responses include a coordinate graph with data plotted correctly.

1 point –
Responses include a coordinate graph with most data plotted correctly, but contains a plotting error (single data point, *y*-intercept or slope).

0 points –
Responses demonstrate no evidence that student has mathematical knowledge of constructing a line graph from a given equation of a line.

Figure 2. Holistic scoring rubric for CR item 6
(Reprinted with permission from the Denver Public Schools)

As indicated by Figure 2, the rubric provides partial-credit for students who plot most of the data correctly, but no credit for students who do not show evidence of being able to translate the equation to the graph. For this item, only ten percent of all students received partial-credit (a score of 1) and only three percent of the entire population of test takers received full credit for this difficult item. As indicated earlier, since one purpose of the test was to provide diagnostic information about students to teachers, each assessment was developed to ensure that items, such as those presented in Figure 1, reflected both difficult and easy content.

Developing Cuts for Evaluating Proficiency

Since the properties of the customized items were largely unknown at the time, assessment staff chose the same raw score cut-points for determining proficiency across all grade levels and content areas in the 2006-07 year. In 2006-07, the following rules reflected the uniform cut-scores based on the percentage of total points correct to determine the proficiency of each student: unsatisfactory students earned 60 percent or fewer points, partially proficient students earned between 61 and 75 percent of total possible points, proficient students earned between 75 percent and 89 percent of total points, and advanced students earned 90 percent or more points. These cuts were selected at the time as a best guess estimate of how students of varying proficiency would perform on these tests.

In the 2007-08 year, district staff recognized that the performance bands needed to be adjusted to account for differences in difficulty found across test forms. The strategy for setting the cut-points involved, first specifying the number of items that the assessment staff wanted to see in each proficiency category, counting the total number of points that would fall under each proficiency category, and then selecting or modifying items to match the performance expectations set by the range of points assigned to each proficiency category. Although the cut-scores varied by content, the cuts remained the same for grades 3 through 8. Figure 3 reflects the set of cuts established for all grades 3 to 8 interim assessments in 2007-08:

Reading Grades 3-8: 32 points - 25 items (20 MC, 3-2pt CR, 2-3pt CR)

Band	Test 1				Tests 2/3			
	Points	Point Range*	% range for band*	% range for half*	Points	Point Range*	% range for band*	% range for half*
U	10	0-10	0%-31%	0%-63%	6	0-6	0%-19%	0%-47%
PP	10	11-20	32%-63%		9	7-15	20%-47%	
P	8	21-28	64%-88%	64%-100%	11	16-26	48%-81%	48%-100%
A	4	29-32	90%-100%		6	27-32	82%-100%	

Math Grades 3-8: 29 points - 23 items (20MC, 1-4pt CR, 1-3pt CR, 1-2pt CR)

Band	Test 1				Tests 2/3			
	Points	Point Range*	% range for band*	% range for half*	Points	Point Range*	% range for band*	% range for half*
U	9	0-9	0%-31%	0-62%	6	0-6	0%-21%	0-48%
PP	9	10-18	32%-62%		8	7-14	22%-48%	
P	7	19-25	63%-86%	63%-100%	9	15-23	49%-79%	49%-100%
A	4	26-29	87%-100%		6	24-29	80%-100%	

Figure 3. Proficiency cut-scores for 2007-08 reading and math interim assessments grade 3-8

As indicated by Figure 3, different performance cuts were set for the easier version 1 test relative to the more difficult versions 2 and 3 of each test. In contrast to 2006-07, the panels in 2007-08 spent more time examining each individual CR and MC item to classify MC items and point values on CR items into different proficiency groups. Since the number of points were established before the items were reviewed, the item panels in 2007-08 were tasked to ensure that there was an adequate number of MC items or points on CR items to correspond with the number of points assigned to each proficiency category. For example, for a given version 1 reading test, if the item panel believed that “unsatisfactory” students could earn one point out of three CR items on that test, then the panel would include eight MC items that could be answered correctly by those students on the test. In this example, including those eight items would ensure that “unsatisfactory” scoring students have the opportunity to meet the 9 point maximum assigned to that category.

Currently and in both years reviewed for this study, proficiency judgments were made based on the overall score, by each standard, and by CSAP framework statement. The proficiency levels for standards and framework statements consisted of two categories of “proficient” or “below proficient”, and 65 percent was set as the threshold for determining whether a student was proficient on a given standard or framework statement. The cut-score set for each standard and for each CSAP framework statement was established by the assessment staff and remained consistent across grades and content area. According to assessment staff, sub-scores and accompanying proficiencies were reported to meet demands from school leaders that the reports should expose weaknesses in teaching standards and framework statements to students. An example of a scoring report reflecting all three levels with scores reported is located in Appendix A-1.

The scoring report presented in Appendix A-1 could be disaggregated to reflect the performance of an individual student on the same information or aggregated to different levels (e.g., by classroom, grade, school, and across schools). Although proficiency is assessed for each framework statement, assessment staff emphasized that district personnel were instructed to focus on student performance at the overall level as well as at the standards level. According to district and school staff interviewed for this study, although the CSAP framework statements provided finer-grained information on student knowledge of each standard, using the statements to drive any kind of decision was discouraged since most statements were measured by only one item. However, it is important to note here that in the 2006-07 year, some teachers compared student performance on the broader standards level for merit pay purposes.

Administering the Assessments

Each school site was provided with dates and guidelines for administering the tests. Prior to each test administration, central assessment staff held regular trainings with the designated school assessment leaders based at each school site to ensure that proper test administration and procedures for scoring were followed. The guidelines and test administration practices closely mirrored the standards and guidelines used when administering the high-stakes CSAP tests with proctors or teachers overseeing the administration of the tests in each classroom. All tests were administered using a paper and # 2 pencil format and accommodations (e.g., providing extra time or reading questions out loud) were provided to students with official documentation demonstrating need to receive accommodations on the test.

The testing period took place at all school sites at the same time during a two-week period. The first testing period took place at the beginning of the year in September, the second period took place at the end of the first semester or beginning of December, and the third test administration took place towards the end of the school year in late April. Following the administration of the test, school sites were given a window of time for scoring all responses and scanning all the scored responses by the deadline issued to all schools by the assessment office.

Scoring the Tests and Scoring Reports

Although the assessment and curriculum staff preferred to have the classroom teacher score the student responses to all items, the composition of raters was determined by each school site. According to central staff, some school sites opted to use teachers instructing in non-core areas (e.g., art, physical education, and music teachers) to score the tests, and in other schools, the science and social studies teachers were recruited for the scoring task. The scoring process typically took place over a two to three day period where teachers were asked to dedicate those days entirely to scoring each test booklet. Due to time and personnel limitations, only one rater was assigned to score all of the responses. The scored sheets were then scanned and entered into a central data repository overseen by a different testing company, referred to as SE in this study, in charge of releasing all scoring reports to schools throughout the district.

After the deadline for scanning all scoring sheets passed, the data were immediately filtered into a secure reporting system created by SE. According to district staff, the process of transferring all data into scoring reports took approximately three to five days. District personnel could easily access these reports through a secured web interface depending on the level of security clearance provided. Queries built into the reporting tool allowed district and school person to aggregate to the following levels: by classroom, across classrooms within a grade level, across all grades within a school, across schools, by region or network within the district, and by the global district view.

Using Data from the Scoring Reports

As noted in the first chapter, the interim assessment data were used by both central district staff and school sites to drive decisions impacting both teachers and students throughout the district. Although this dissertation study focuses on district-wide uses as mandated by central staff members, assessment staff and teachers interviewed in this study noted that schools also chose to use these assessments as determined by school leaders. Examples of site-specific uses include: leveling students based on the first test to determine ability group assignment for each semester; enacting instructional changes such as focusing on particular standards across classrooms; and, using the data to identify whether students belong to the second or third tier in a response to intervention (RTI) structure. Regarding the last use, a set of interventions are designed by each school to ensure that students falling behind grade expectations receive appropriate levels of support.

Currently, a few of the items purchased from TPR populate the interim assessments used in grades 3-8 and item panels are no longer convened since the district has been using the same test forms for those grades since the 2008-09 years. However, the same number of items and the same approach to determining cut-points for proficiency in 2007-08 still apply to this date. Despite this study’s focus on the 2006-07 and 2007-08 years, the findings in this study still apply to the case study district and could inform their continuing efforts to improve and refine their interim assessments and could likewise inform other school districts across the county seeking to use interim assessments to inform multiple decisions.

Comparing the Test Development Process with the CSAP

Although the interim tests used in this district were created to measure the same content areas and standards largely tested by the CSAP, the process described in the previous section to develop the interim assessments differs substantially from their higher-stakes state (CSAP) counterpart. That is, it is important to keep in mind that some of the comparisons being made between the interim assessments and the CSAP in the following chapters may vary due to the different approaches and timeframe used to develop these two test systems. Since state tests are designed specifically for the purpose of meeting high-stakes accountability mandates as defined by both federal and state laws, these test would typically require more stringent procedures and checks for test development. For example, the steps outlined below reflect the test development process adhered to by the developer of the CSAP tests, CTB Mc-Graw Hill[4], when developing tests for either state or local (district) clients.

Develop Assessment Materials

- CTB content editors—former teachers, curriculum experts, and trained test development professionals—research and collect materials, write items, and develop scoring rubrics.
- Artists, designers, and writers work together to ensure graphic and textual consistency.
- Verification processes are established and strictly followed. These processes include reviews for curriculum match, grade-level appropriateness, equity, and graphic and textual coherence.
- Information systems specialists design score reports to align with test specifications and objectives.

Review Assessment Materials

- CTB test development specialists review all materials to ensure that test specifications and objectives are met. Editors review materials for clarity, effective item construction, and balanced representation of ethnic, gender, age, and role images.

Conduct Pilot Tests and Usability Studies

- Assessment materials and score reports undergo extensive classroom pilot testing and usability analysis to ensure easy use by students and educators.
- Special consultants, including specialists in early childhood and limited English proficiency (LEP), review materials, and provide feedback about content and item formats.

[4] Referred to as CTB from this point onwards.

Conduct Tryout Studies

- Assessment Research staff conducts a tryout of assessment materials, using a carefully selected and appropriate sample (a national sample for a national standardized test or a state sample for a state assessment). Item analysis is then completed.
- Scorers review constructed responses and rubrics to ensure consistent and accurate scoring.
- Test development and research staff reviews statistical data and the comments of teachers and students. This information guides the revision process.

Conduct Bias Reviews

- Teachers, content specialists, sensitivity specialists, parents, and LEP representatives review test materials and provide feedback about content validity and equity.

Conduct Pilot Test

- In some cases, particularly in state programs, full-blown implementation is often preceded by an interim assessment in which the test is piloted, while research and development continues. In either case—interim assessment or full implementation—districts and states must take care to carefully monitor the process.

Produce Standardization Materials

- Research staff and content specialists select final test items, using CTB's computer-based item selection program to optimize statistical properties and content coverage.
- Designers and usability analysts review final page design and item layout to ensure they represent student performance.
- Content specialists and editors conduct an intensive review for content validity and editorial accuracy
- Customized state tests are typically designed to assess pupil proficiency in meeting specific state curriculum standards. In many cases, however, states also desire information comparing their students to those in other states throughout the nation. In such instances, previously normed test items or a short survey version of a norm-referenced test may be embedded in the state assessment. This enables both standards-based and norm-referenced information to be generated.

Produce Final Materials

- Manufacturing staff produces final assessment materials.
- Assessment report designers develop final reports that teachers and administrators can use for instructional improvement.
- Test development staff provides comprehensive support materials and activities designed to assist students, parents, teachers, and school administrators before, during, and after the test

In contrast to the compressed duration of time and resources dedicated by district staff to develop each set of interim assessments used in DPS, the process described above by CTB to develop either state or customized assessments would require a period of a few years (for state level tests) or a few months (for customized assessments) before the test becomes fully operational at the state or district-wide level. That is, as outlined above, a test developed by CTB undergoes several rounds of content review and analyses, field testing and refinement before being considered by the test company as an instrument that can reliably assess student performance. In addition, as described above, scorers receive training to ensure consistent and accurate scoring and statistical data are reviewed to help guide the panel members and refine the test. This iterative process described by CTB to create their test instruments closely parallels the process utilized by other test vendors seeking to develop reliable tests to meet high-stakes state and federal mandates, but does not necessarily reflect the standards or process used by all test vendors and other institutions developing interim tests with varying levels of stakes involved.

As described in the first chapter, since there are no uniform standards used by various institutions developing interim tests, the process of developing these test systems could look similar to the process jointly established by TPR and DPS, or could undergo the longer time horizon of field testing, piloting and refinement prior to reaching the operational phase as described by CTB. Further, compared to the process pursued by DPS, the development process described by CTB would require a substantially larger investment from a district or state entity than a process that is largely directed by local personnel. In other words, not all districts may afford purchasing the products created by vendors such as CTB and may subsequently select alternative or more affordable options such as the one offered by TPR where district staff are largely in charge of steering the test development process. Regardless of how the test development process is structured in a given school district, Shepard's (2009) recommendation to find supportive evidence for using interim tests to fulfill different purposes and claims applies to all interim assessments. Due to the large amount of resources being invested by school and state entities in interim programs, and the recent developments at the Federal and state level to employ interim tests for evaluating teacher performance, additional scrutiny is warranted to evaluate these interim assessment products.

CHAPTER 3 - Utilizing an Argument-Based Approach for Validating Tests

According to the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education's (NCME) *Standards for Educational and Psychological Testing*[5] (1999), "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests…The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. When test scores are used or interpreted in more than one way, intended interpretation must be validated" (pg. 9). To accumulate evidence supporting the interpretations or uses of a test, the *Standards* recommends that researchers develop an argument or a list of propositions supporting a given test use and then design a set of studies that address the argument or propositions. The recommendation to conduct a validity study through an argument-based approach dates back to earlier works by Cronbach (1980, 1988), Kane (1992), and Shepard (1993). Although validity theory has evolved over many decades, Shepard notes that the pace of change accelerated in the 1980s when conceptualizing validity studies as an argument emerged as an organizing theory for evaluating test interpretation and uses.

Prior to the momentum gained in the 1980s toward using an evaluation argument based approach to evaluate test validity, researchers typically focused validation efforts on gathering evidence for either the content, construct, or the criterion-oriented aspects of a test. Shepard (1993) points out that this segmented "trinitarian" approach allowed researchers to claim that a test was valid by simply providing evidence addressing one of these three aspects of a test system. Cronbach and Meehl (1955) explicitly rejected this trinitarian approach by establishing construct validity or test use as the overarching focus of any validity study. According to Cronbach and Meehl (1955), "the investigation of a test's construct validity is not essentially different from the general scientific procedures for developing and confirming theories" and studies addressing just the content or criterion-oriented aspects of a test are insufficient (pg. 302). Cronbach (1981, 1988) later modified this focus on construct validity by proposing that construct

[5] Referred to from this point onwards as the *Standards*.

validity be studied and framed as an evaluation or interpretive argument where the most relevant questions are prioritized and subsequently evaluated to validate score interpretations and uses. Although this concept of validating test score interpretations and uses through an evaluation lens is also conveyed in Messick's (1989) seminal work "Validity", Shepard (1993) notes that his highly complex and extensive treatment of evaluating the multiple facets of construct validity, including making explicit consequences and value-laden judgments relative to test interpretation and use, makes it difficult for researchers to understand how best to prioritize "which validity questions are essential to support a test use". Further, Shepard states that Messick's pronouncement "that construct validation is a never-ending process" appeared to give "practitioners permission to stop [their validity investigations] with incomplete and unevaluated data" (pg. 407).

To help practitioners think about ways to organize and conduct a validity study drawing from the multiple facets of construct validity laid out by Messick, Shepard points to Cronbach's (1998, 1999) and Kane's (1992) use of an evaluation argument approach[6]. Under an evaluation argument approach, a set of clear questions or assumptions are first established to examine claims about what a test can do and then a set of studies are designed to gather empirical evidence to directly address those questions and assumptions. According to Shepard (1993), because it is impossible to evaluate "the whole program in a comparatively short period of time" and to do this exhaustively, evaluation design involves identifying the most relevant questions and deciding what emphasis should be given to each" (pg. 430). Moreover, Shepard notes that since test use serves as the central locus for organizing a validity study, if two uses of a test differ in substance, then different lines of investigations would have to be pursued to examine these two distinct uses (pg. 430).

The guidelines laid out in the *Standards* takes the concluding recommendations of Shepard (1993) by encouraging researchers to evaluate a test based on an evaluation argument approach, but stops

---

[6] Shepard (1993) uses the term "evaluation argument" to define the process of establishing the argument to evaluate a given test use and designing the studies to support the argument. In Kane's "Validation", the evaluation argument described by Shepard is separated into two separate components: the interpretive argument and the validity argument. The interpretive argument refers to the argument established to evaluate a test use and the validity argument refers to the set of studies designed to evaluate the interpretive argument.

short of her recommendation to provide researchers with a framework for operationalizing this approach. The *Standards* identifies five key sources for evidence[7], taken directly from the same areas covered earlier by Messick (1989), to gather in a validity study and states, "The decision about what types of evidence are important for validation in each instance can be clarified by developing a set of propositions that support the proposed interpretation for a particular purpose of testing" (pg. 9). Aside from recommending a set of propositions, the *Standards* does not provide an explicit framework for helping researchers understand how many propositions are needed and what criteria should be used to organize the "few lines of solid evidence" supporting a particular proposition.

Although the suggestion to develop propositions and organize studies examining those propositions is consistent with an evaluation argument framework, Chapelle, Enright and Jamieson (2010) note that the process of devising a list of propositions addressing each of the five sources can lead to an "endless list of possibilities" for investigating test validity. That is, the lack of specific guidelines for framing a test validity study in the *Standards* makes it unclear as to how one should organize or proceed with such a study. Echoing the same recommendation raised by Shepard more than a decade ago, Chapelle et al (2010) point to an evaluation argument based approach, specifically, Kane's (2006) interpretive and validity argument framework as a way to prioritize, organize and conduct a validity study by drawing evidence from the five sources highlighted in the *Standards.*

Using an evaluation argument based framework as recommended by Cronbach (1980, 1999), Shepard (1993), and Kane (1992, 2006) serves as the basis for evaluating the uses of the interim assessments in DPS for this dissertation study. More specifically, this study takes the recommendation in Kane's (2006) "Validation" to undertake a two-step process of evaluating test validity by first establishing the network of inferences and assumptions supporting the specific use of a test, and then undertaking the empirical analyses to test out the established network of supportive inferences and assumptions. According to Kane (2006), the former step represents the "interpretive argument" which "makes the

[7] According to the standards, these five sources are: evidence based on test content, evidence based on response process, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing.

reasoning inherent in the proposed interpretations and uses explicit so that it can be better understood and evaluated", and the latter step represents the "validity argument" which serve as the "evaluation of the interpretive argument" (pg. 23).

After establishing the interpretive argument or the network of inferences and assumptions supporting test use, the first phase of the validity study entails evaluating whether the logic developed to support test score interpretations and uses are deemed clear, coherent and plausible (pg. 29). The second phase of the validity study entails designing studies that test out the interpretive argument. In "Validation", Kane provides several examples of how the interpretive and validity argument could look for different types of test use and points to many of the same five areas highlighted in the *Standards* as appropriate sources for finding supportive evidence for the evaluation of test use. In one of Kane's examples, evaluating a test used for making placement decisions, the framework for evaluating this type of test use could encompass assessing aspects such as: the scoring rules and guidelines set to score student performance on items; the reliability and accuracy of the assessment in determining the performance level of a student; and the extent to which the findings from looking at all prior studies support the use of the test for determining the set of courses that a student should take in order for her to reach competency. Before the validity argument or evaluation of the interpretive argument ensues, the inferences (scoring, reliability or generalizability, extrapolation and decision) and underlying assumptions supporting each inference (e.g., the assumptions that raters are scoring students consistently and using well-established scoring guidelines and rubrics to appraise performance on tasks) are first examined to determine whether they clearly support the specified test use. If the framework established for pursuing an evaluation of test strikes readers as clear, coherent and reasonable, then studies designed to draw from areas identified by the *Standards* such as examining the response-process to provide evidence for scoring or examining the test relative to an external criterion to support the extrapolation inference, could be conducted to provide supportive evidence for test use.

Kane, like Shepard (1993), notes that the design of the validity study as expressed in the interpretive argument used needs to be responsive to different uses. For example, if the same test used for placement purposes is also used to help inform what teachers should do with individual students within a classroom context, then a new interpretive argument would be constructed to examine how effectively teachers can "integrate data from a variety of sources, most of which are not standardized, to develop a holistic view, or model, of each student" (pg. 46). Under a qualitative interpretation of test use, the interpretive argument may not require drawing evidence from four of the five areas identified in the *Standards* calling for largely technical or quantitative methods for evaluating groups of students. However, the interpretive argument would ideally include developing an argument to evaluate the fifth area, examining test consequences, identified in the *Standards* to determine the benefits or impact of this type of assessment on an individual student.

Although Kane's argument-based approach forms the basis for the conceptual framework used in this dissertation study, this study modifies his argument to ensure that the inferences and assumptions conform to the specific uses and context being examined in this study, and to ensure that a broad audience readily understands the argument. To provide specific examples of modifications made, the interpretive argument Kane uses to make trait-inferences about students is explained within the context of the first three uses that require obtaining a reliable measure of student proficiency over the content being measured by each test. In "Validation", Kane presents several interpretive arguments to address different uses or purposes for a test, but presents the same sequence of three inferences (scoring, generalization and extrapolation) in each argument with many of the same assumptions repeated under each inference. These three inferences lead to either an "implications" or "decisions" inference which assumes that if all prior inferences hold, then the data from the test could be used to drive or support particular decisions and those uses are largely beneficial or harmless. Since the first three uses of the test examined in this study and the stakes associated with each use depend largely on the proficiency level attained by an individual student on each test, Kane's argument for classifying a respondent's trait level presented in "Validation" serves as a reasonable model for crafting an interpretive argument for each of the first three uses. Figure 4 presents a diagram of Kane's interpretive argument for making trait inferences, beginning from the trait or characteristic of interest inferred from the test, to "implications" as defined by how information gained about the trait is operational within the context of a specific use.

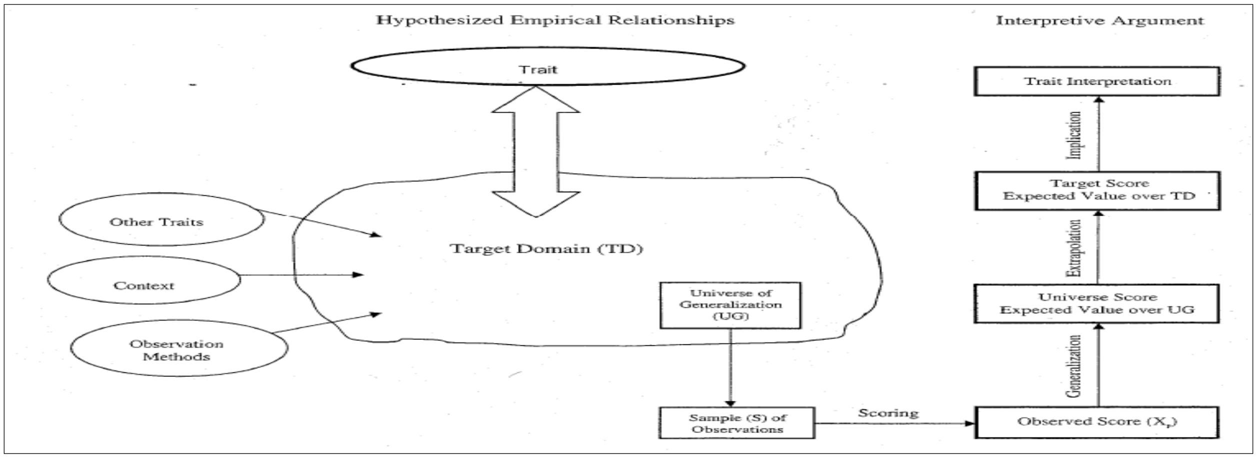


Figure 4. Kane's measurement procedure and interpretive argument for a trait-interpretation (from Kane, 2006)

To explain each step of Kane's interpretive argument and where this study modifies or departs from Kane's framework, the argument in Figure 4 is explained within the context of the grade 8 math, 2006-07, version 2 test used to fulfill three out of the four uses of the interim assessments evaluated in this study. This test and its contents are analyzed and discussed further in the next chapter.

According to Kane, "A trait is a disposition to behave or to perform in response to some kinds of stimuli or tasks, under some range of circumstances" (pg. 30). Although the different multiple-choice (MC) and constructed-response (CR) items on the grade 8 math test serve as the stimuli or the tasks within the context of this study, the test was not explicitly designed with the intention to measure an underlying disposition but was designed to measure what a student already knows at the time the student is tested. In other words, the score earned on the grade 8 math test simply determines a student's assignment into one of four proficiency levels on the grade 8 math test. Since there is no specified trait being measured by the grade 8 math test, the argument used in this study removes any references about a trait, and evaluates the interpretive argument relative to a student's mathematics proficiency (the "Trait", "Trait Interpretation" and "Other Traits" components in Figure 4) as presented below in Figure 5.

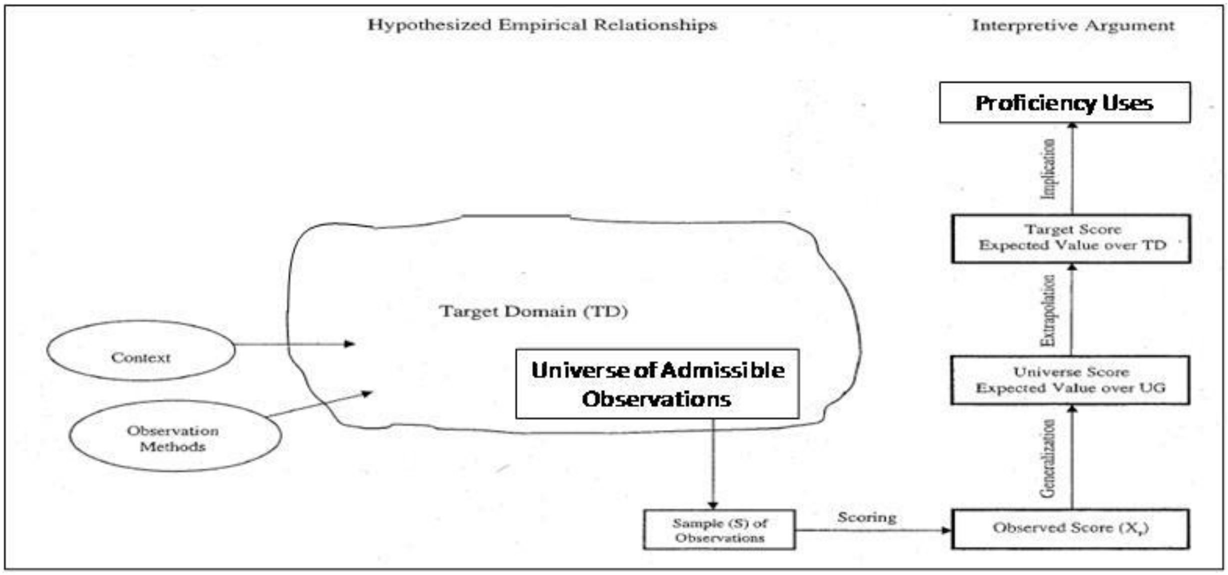


Figure 5. Modified measurement procedure and interpretive argument for a domain-based interpretation

The Target Domain in Figure 5 represents the set of critical grade 8 math skills and knowledge defined by the Colorado Department of Education (CDE) as "math literacy skills and knowledge needed for citizenship and employment in the 21s century" (CDE, 2005). More specifically, these critical skills and knowledge represent the state's grade 8 standards and benchmarks for math. There are six math standards that represent the general body of knowledge required of all students in every grade and the benchmarks represent the specific grade level expectations for a given standard. The six math standards are described by CDE as follows:

Standard 1
Students develop number sense and use numbers and number relationships in problem-solving situations and communicate the reasoning used in solving these problems.

Standard 2
Students use algebraic methods to explore, model, and describe patterns and functions involving numbers, shapes, data, and graphs in problem-solving situations and communicate the reasoning used in solving these problems.

Standard 3
Students use data collection and analysis, statistics, and probability in problem-solving situations and communicate the reasoning used in solving these problems.

Standard 4
Students use geometric concepts, properties, and relationships in problem-solving situations and communicate the reasoning used in solving these problems.
Standard 5
Students use a variety of tools and techniques to measure, apply the results in problem-solving situations, and communicate the reasoning used in solving these problems.

Standard 6
Students link concepts and procedures as they develop and use computational techniques, including estimation, mental arithmetic, paper-and-pencil, calculators, and computers, in problem solving situations and communicate the reasoning used in solving these problems.

For each standard, grade level expectations are further delineated through benchmarks. For example, under Standard 2, one benchmark reflecting grades 5-8 expectations states that a student in these grades should be able to "distinguish between linear and nonlinear functions through formal investigations". Appendix B-1 provides information on all of the math benchmarks that apply to grade 8 math. Kane's depiction of the target domain in Figure 4 has no specific shape since the target domain can be "defined very broadly (e.g., Caroll's definition of intelligence); others are more circumscribed but still broad (e.g., proficiency in algebra), and some are quite narrow (e.g., skill on a specific task)" (pg. 30).

The two ovals at the left of the target domain labeled "context" and "observation methods", represent other ways outside of the formal testing context (e.g., taking a standardized test during a constrained time period on a designated school day) and method (e.g., answering multiple-choice and constructed-response questions) used by the grade 8 math interim test to assess a student's ability on the target domain with respect to the grade level skills and knowledge defined by CDE. One example of a different context would be having a student participate in a project based performance task that requires her to apply math skills. The student could be asked to play the role of a government advisor on poverty issues and write up a report which includes having to graph frequencies and present statistics on poverty rates (aligned with Standard 3). For this project-based example, the observation method consists of the student's work as reflected by the final report. Another example would be a math computer program such as Cognitive Tutor, used in many high schools and a few middle school classrooms across DPS, that engages students in interactive lessons. The computer based platform serves as another context for observing a student's performance on the target domain. The observation method for the computer based example consists of the student's responses to the different levels of complex problems provided by the software program. An example of a context outside of the school setting would be having the student account for daily cash flow associated with a store's inventory in work environment. In this work context, the observation method is the student's application of skills addressing both Standard 1 (number sense) and Standard 6 (applying computational skills) to track, report, and document revenue flows and inventory in the store.

In the diagram represented in Kane's figure (Figure 4), the box in the Target Domain labeled "Universe of Generalization" represents a subset of the target domain. The universe of generalization represents the formal testing context within each classroom and the set of MC and CR items selected to measure students on grade 8 math standards and content. The universe of generalization is a term used in Generalizability Theory or G-Theory (Cronbach, Gleser, Nanda and Rajaratnam, 1972; Shavelson & Webb, 1991; Brennan, 1992). In the modified diagram shown in Figure 5, the universe of generalization is replaced by the universe of admissible observations since in G-Theory, a student's observed score comes from a "universe of admissible observations, observations that a decision maker is willing to treat as interchangeable for the purposes of making a decision" (Shavelson & Webb, pg. 3). This means that a student's score on the grade 8 math interim test is derived from a sample of any possible CR and MC item reflecting grade level content and standards, any acceptable rater that could have been used to evaluate a student's score, or any possible occasion or time point that could have been used to administer the test to the student. The universe of generalization is directly related to the "universe score" that is addressed later in the interpretive argument under the generalization inference. A brief overview of G-Theory is presented here and related back to the context of the grade 8 math interim test to facilitate understanding of both the universe of admissible observations and the universe of generalization.

Shavelson and Webb (1991) describe G-theory as "a statistical theory for evaluating the dependability ("reliability") of behavioral measurements". In classical test theory, a student's observed score on a given item consists of a "true score" and error, or X (observed score) = T (true score) + e (error). The error term (e) represents the concept in classical test theory that an instrument cannot measure observed performance perfectly and that there will always be a degree of error associated with any measurement. G-Theory extends classical test theory by identifying different sources or facets within the testing situation that could potentially increase the degree of error associated with a student's performance. These facets could include: the number of raters used, the number of items on a test, the number of test forms administered and the number of occasions or times that tests are administered. The error left over or not accounted for by the specified facets is considered to be random or "noise" caused by other possible sources outside of the testing situation that cannot be directly estimated or controlled. In G-Theory, specifying all known facets (e.g., raters, items, or occasions) that may contribute to obscuring the true performance of a student constitutes the first step of a generalization-study (*g*-study). Each distinct facet in a testing situation that could directly affect a student's "true score" comprises the "universe of admissible observations" or the box located in the Target Domain of Figure 2.

The second step of a g-study entails estimating the variance components associated with each specified facet in the universe of admissible observations. In this dissertation study, since only one rater is used to score a given test and only one occasion is used to administer each distinct interim test, measurement error can only be estimated for one facet (the items) relative to the object of measurement (students taking the test). A single-facet universe means that any CR or MC item in the entire pool of possible CR and MC items measuring grade 8 math content as defined by the state standards benchmarks represents the universe of admissible observations. If data were available to assess raters, or if the test was administered to students over different occasions or time points, then the variance components associated with all possible combinations of each rater, time point or occasion, and item would be estimated for the grade 8 math interim test.

Turning back to the diagram in Figure 5, the box labeled "Sample of Observations" located under the universe of admissible observations box represents the specific set of MC and CR items that were used on the grade 8 math interim test. Moving to the right of the "Sample of Observations" in Figure 5, the scored responses from taking the test based on the sample of items used to measure grade 8 math knowledge and skills, represents the “observed score”. In “Validation”, the first inference of the interpretive argument begins at this point in the diagram (the scoring inference). However, in this dissertation study, the interpretive argument begins earlier with two inferences acknowledged prior to the scoring inference: content design and item design. These two inferences are added to the interpretive argument used in this study to acknowledge the importance of ensuring that adequate work was undertaken to ensure that the blueprints and content reflected on the test adequately represented key standards and benchmarks (content design inference) and each test consists of a variety of test items selected to measure students of varying proficiency (item design inference).

Moving up from the “observed score” in Figure 5 is the “universe score”. The universe score represents a score derived from a constrained version of the target domain. Since decision makers often times use information from a test to generalize about a student’s performance, the universe of generalization then, refers to the conditions of a facet to which a decision maker wants to generalize (Shavelson and Webb, 1991). In G-Theory, a decision-study or a *d*-study is used to estimate the standard error of measurement (SEM) and a reliability coefficient over a fixed number of conditions drawn from the universe of admissible observations. Whereas a g-study is used to estimate the error associated with a student’s score relative to a given item, rater, test form or occasion, a *d*-study evaluates the extent to which a student’s observed score is reliable and observed over specific conditions for each facet specified earlier in the *g*-study. For example, a decision-maker may want to generalize a student’s performance over 30 items, three raters and three test occasions. In a *d*-study, the reliability of that student’s score is then evaluated relative to the specified conditions (in this case, 30 items, three raters and three test occasions) set for each facet. Applying a *d*-study for the grade 8 math test, would answer questions about whether all possible sets of 18 items used out of the entire pool of acceptable items could drive decisions for each of the uses evaluated in this study. These decisions could be criterion based (e.g., whether a student located below a specified threshold requires extra tutoring) or normative based (e.g., an award goes to the top-twenty ranked students in a school).

After establishing the condition (the universe of generalization) to which a decision maker wants to generalize, the universe score "is defined as the expected value of his or her observed scores over all observations in the universe of generalization (analogous to a person's "true score" in classical test theory)" (Webb & Shavelson, 2005, pg. 92). A universe score can be understood within the context of a thought-experiment where repeated measurements or observations of a student are taken using interchangeable conditions specified in the *d*-study. This would mean that a decision maker using the grade 8 math test to drive decisions, makes generalizations about a student's performance over all possible sets of 18 CR and MC items that could have been used to evaluate a student's math knowledge and skills. If a different set of CR and MC items were used to evaluate a student on the same grade 8 math test, a student's observed score may differ from the observed score earned on the first set of items used to measure a student's grade 8 math performance. Under G-Theory, if repeated measurements of a student were taken on all admissible sets of CR and MC items used to measure grade 8 math skills and knowledge, the average score over all those measurements would represent the universe score.

However, in the absence of being able to take repeated measurements or observations of students on interchangeable sets of items, the precision of a student's observed score is evaluated. That is, if a student's observed score could be measured with reasonable precision, this finding would provide some supportive evidence that a test user may infer that a student's performance on these items could be similar to the same student's performance on a different set of acceptable or admissible MC and CR grade 8 math items. Based on this logic, the decision maker would be able to use the information from the one test to support specific uses of the test such as, in the case of the grade 8 math test, identifying which students require mandatory remediation prior to entering high school.

In this study, any term more commonly used in G-Theory, such as "universe score" or "universe of generalization" is reworded or reframed to make the language used in the interpretive argument more accessible to a broader audience. For example, an assumption supporting the generalization inference for this study states, "a student's performance on a selected sample of all possible MC and CR items used on the interim test, is estimated with reasonable precision". Although the term universe score is not mentioned in this assumption, the assumption preserves the sampling connection between the set of items reflected on the test relative to the set of all possible items that could have been used to measure grade 8 students on grade level content and benchmarks.

In addition to avoiding the use of specialized terminology in the interpretive argument, another reason for not using the g-theory language in the interpretive argument is that those terms directly imply that G-theory or *g*- and *d*-studies will be used as part of the validity argument. To remain consistent with the G-theory language used in the interpretive argument, Kane suggests using *g*- and *d*-studies to investigate whether a test produces dependable or reliable test scores under the generalization inference. In evaluating the properties of the grade 8 math test, because the only facet that can be evaluated in the universe of admissible observations is items, the reliability estimate produced under a d-study with a single-facet design would yield a similar estimate to one derived under classical test theory. In other words, not much more information is gained from using a *g*- or *d*-study when only one facet is specified. An important consideration to note relative to findings presented for the generalization inference in the next chapter is that without being able to factor in the impact of other facets in the universe of generalization on a student's score, the reliability coefficient estimate will most likely be biased upwards. That is, a more optimistic or higher reliability coefficient will be generated when specifying only one facet, since the variability in scores associated with other facets known to have an impact on observed performance, such as raters is not accounted for in the study design.

After establishing whether the universe score is reliable under generalization, the next piece of the interpretive argument represented in Figure 5 is the extrapolation from the Universe Score to the "Target Score" or the expected value over the Target Domain being measured by the test. Unlike the universe score, the target score moves beyond the restricted universe or context of all possible CR and MC items that could be used to measure students, and locates the student within the Target Domain of interest. Within the specific context of this study, this target score means that a student who is considered to be proficient under the UG should also be considered as proficient by any other contexts and methods that could have been used to assess a student's proficiency on grade 8 math skills given unlimited time and resources. For example, a student classified as “proficient” on the grade 8 math test should also be classified as proficient when the context changes from a formal testing situation to a real-world context (e.g., the inventory example noted earlier) and the observation method changes from responding to formal test items to completing a math project that corresponds to a specific standard. In another example, a teacher’s appraisal of the same proficient student’s response to her math question would confirm that the student appears to have sufficient knowledge on a particular benchmark. Since the target domain represented by the CDE standards and benchmarks are broad, and the grade 8 math interim test limits the number and types of tasks (CR and MC items) to measure students, it is not possible, with the data sets available to evaluate the interpretive argument or to gather sufficient evidence demonstrating that a single math test with only 18 items provides sufficient information about a student’s performance over the target domain. Although scores from the interim test are compared with scores from another formal testing context and method (the larger end of year summative test) addressing the state standards and benchmarks, a thorough investigation of this inference would include having to evaluate other data points not available such as comparing how students perform on the grade 8 math test relative to their performance as reflected by student portfolios containing work completed in the classroom and evaluated by a teacher. In addition, evaluating the extent to which the interim test adequately captures the content standards and benchmarks would require enlisting curriculum specialists to evaluate the alignment and depth of content reflected by each test relative to the state standards and benchmarks.

In Kane’s original figure of the interpretive argument (Figure 4), “Trait Implications” immediately follows the target score under the extrapolation inference. In the modified diagram presented in Figure 5, Trait Implications is replaced by “Proficiency Uses” in order to address the concrete or specific ways in which the district has used the data based largely on a student’s proficiency to inform each of the three areas evaluated in this study. For the grade 8 math test, this inference addresses the use of the proficiency data to identify unsatisfactory performing students for mandatory summer remediation, to determine whether teachers are meeting their student growth objectives to earn a bonus under the teacher compensation program, and to gain predictive information for both planning and intervention purposes. This inference not only addresses how the proficiency data are used but would also consider the extent to which these uses are beneficial. The next section of this chapter presents the interpretive argument and assumptions used to support all four uses of the interim tests examined in this study. In the following section, one argument is presented for the first three uses that rely largely on the reliability and the accuracy of the scores to determine student proficiency over state standards and benchmarks, and another argument is presented for the fourth use, which depends largely on the extent to which teachers are able to use the interim test data to improve their understanding of their students and their own instructional practices.

Interpretive Argument for Evaluating Interim Assessments

By drawing on Kane's interpretive argument framework as the basis for evaluating the interim assessments for this dissertation study, this section presents the argument addressing each of the four uses. To reiterate, the four uses of the interim tests evaluated in this study are:

1) Predicting the performance of students on the high-stakes end of year state assessment to identify which students need additional remediation or tutoring before the testing window and to plan for future interventions and restructuring in classrooms or in schools.
2) Identifying unsatisfactory students for mandatory remediation.
3) Evaluating and rewarding teachers on the basis of growth made by students between the first and third interim test administration.
4) Changing or adjusting instructional practices based on diagnosing student misconceptions with content from the interim tests.

As discussed in the previous section of this chapter, the first three uses specified above depend largely on the proficiency level attained by a student on the set of items tapping into the target domain measured by each test. The target domain for each test represents critical grade level skills and knowledge in math, reading, or writing. These skills are represented by the state standards and accompanying benchmarks reflected in the blueprints and the scoring reports of the interim assessment program. According to the Colorado Department of Education (CDE), the state standards and benchmarks represent the set of math, reading or writing knowledge and skills that students need to master in order to become successful in a post-secondary environment. Since the actions under the first three uses are triggered after establishing a student's proficiency, the inferences and assumptions from content design to extrapolation are relevant to the interpretive arguments for all three uses. The interpretive argument for each distinct use diverges after extrapolation to address the specific context pertaining to each use.

The fourth use evaluated in this study evaluates the instructional use of the interim tests for making changes to teaching practices. More specifically, this study evaluates the extent to which the data from these interim assessments can help inform a given teacher's overall contextual framework of what each student knows and can do, and the extent to which teachers use the data from interim assessments to adjust or change their teaching practices to meet individual learning needs. For this fourth use, a different set of inferences and assumptions were used in this study to evaluate this claim that these tests can help teachers improve their instruction to meet individual student needs.

Figure 6 in the following page presents the set of inferences and assumptions that apply to the first three uses evaluated in this study. There is an implicit assumption being made by DPS staff that these interim assessments provides reliable information about student performance on content assessed since two of the three uses (mandatory remediation and teachers who opted to base their student growth objectives using only data from interim assessments) were driven solely by data from these tests. This implicit assumption about being able to generalize a student's performance beyond the set of test items and outside of the formal testing context is not unique to this district. Decision-makers at other district, state and federal levels have also encouraged the use of formal summative and interim assessments to drive high stakes decisions. For example, in Colorado, the state's former policy to convert public traditional schools maintaining an "unsatisfactory" designation for three consecutive years to a charter school with entirely new management and staff was driven entirely by decisions from one summative state test program.

For the first inference in Figure 6, content design, the assumptions reflect the expectation that the content reflected on each blueprint and test is aligned to the state standards emphasized in the CSAP. Since the interim assessments are designed to measure the same set of key standards, the second assumption supporting content design reflects the expectation that the blueprints and test items also reflect content and similar tasks used by the CSAP. The district's efforts to align the interim assessments with key standards that are heavily weighted on the CSAP tests were done to help fulfill two district wide objectives. One objective was to design interim tests that would predict the future performance of students on the CSAP and the other objective was to give teachers useful data on how well students have mastered state standards during the school year.

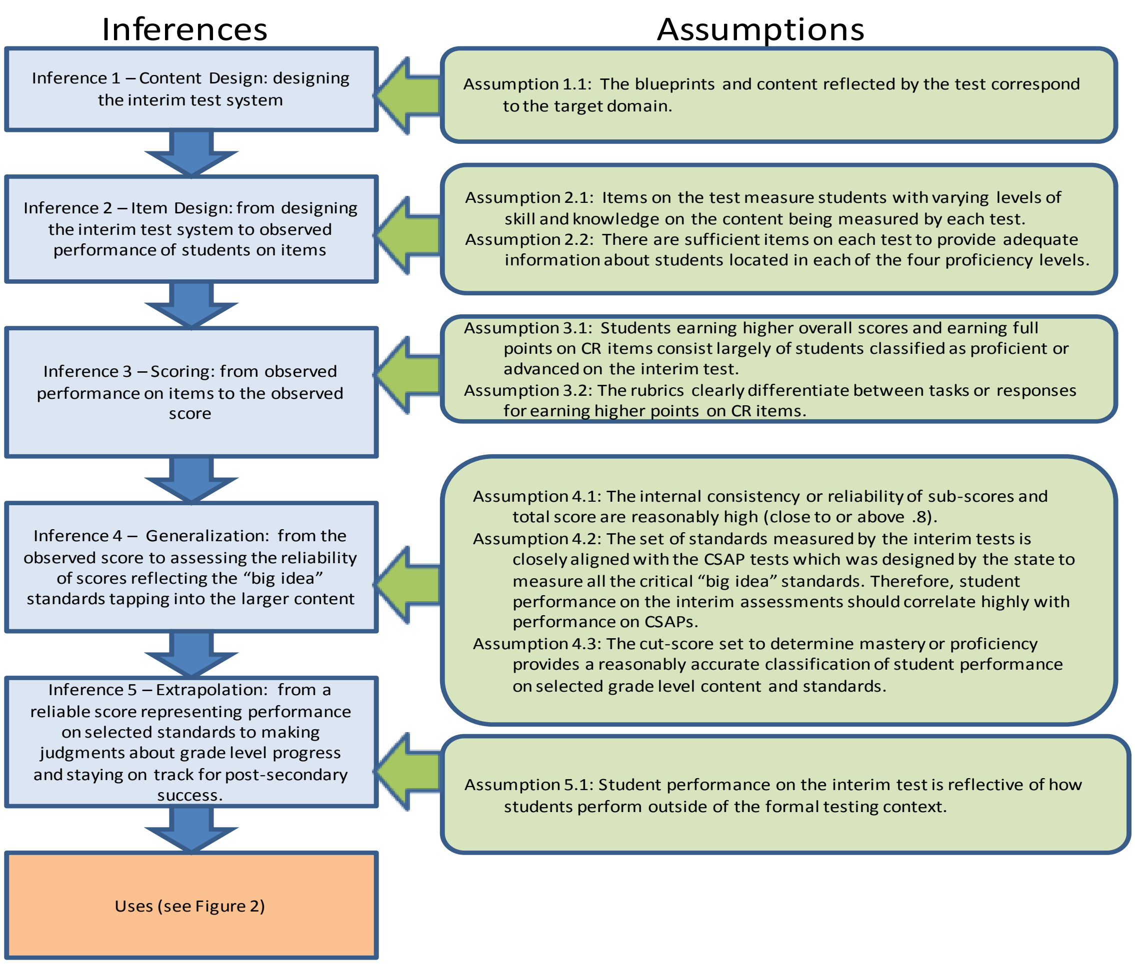


Figure 6. Arguments supporting inferences from Content Design to Extrapolation

The item design inference and supporting assumptions in Figure 6 builds on the content design inference. That is, if the blueprints and the interim tests were designed to provide information on student performance relative to the standards and the CSAP content, then the items should be designed to differentiate between students located at varying levels of proficiency. As noted in the first section of this chapter, the assumptions supporting the item design inference in Figure 6 reflect the expectation that the test consists of items that measure students with varying proficiency over the content represented on each interim test. This inference also assumes that there should be a few items on each test measuring students at different proficiency levels in order to gain adequate information about different types of students at each proficiency level.

The assumptions supporting the scoring inference build on the same set of expectations supporting the item design inference. That is, since the items were designed to differentiate between students of varying proficiency levels, the scored responses should conform to that expectation where higher achieving students are given higher scores for providing higher level responses to CR items and for endorsing more MC items including more difficult MC items. The assumptions supporting the scoring inference reflect the expectation that the characteristics of higher-order thinking is embodied in each scoring rubric and that the scores of students should reflect clear differences between lower and higher performing students across items.

The assumptions supporting the fourth inference shown in Figure 6, generalization, determines the extent to which the instrument yields reliable information about students based on the set of MC and CR items selected out of all possible or acceptable MC and CR items that could have been used to measure students in each grade and content area. Supportive evidence checking the assumptions under generalization include establishing whether the instrument appears to meet reliability standards and ensuring that students are measured with reasonable precision. Reasonable precision in this case means that students located close to each proficiency category threshold could potentially be estimated as falling under two proficiency categories due to a small degree of error associated with their scores. However, if the scores of many students located farther away from the established proficiency cuts are measured with considerable error, then this finding would call into question whether the sample of CR and MC items used on an interim test can measure a student's score with reasonable precision.

Whereas generalization makes assumptions about student performance relative to the set of CR and MC items represented on each test, extrapolation makes assumptions about a student's performance within the larger context of the target domain being measured. The target domain for the interim assessments evaluated in this study is bounded by the state's grade level and content standards and benchmarks measured by each test. For example for the grade 8 math interim assessment addressed in the earlier section, the MC and CR items administered within a formal testing context tap into the set of skills defined by CDE as required for post-secondary success either within a tertiary or a work setting. If one believes that the items effectively tap into the target domain, then a student's score from the interim assessments not only provides a good signal for whether a student has acquired proficiency over the target domain, but as discussed in the previous section, this score should also be aligned with other measures or observations measuring a student's performance over the target domain. Evidence to uphold this assumption about this connection between an interim test and the target domain would include showing the extent to which the items on the test addresses key Colorado state standards and benchmarks, and demonstrating the extent to which a student's performance on the interim test is closely aligned with performance on measures using either a similar (e.g., the CR and MC items used in the formal testing context of the CSAP) or different context (e.g., the Cognitive Tutor software program used in classrooms) for measuring a student over the target domain.

Figure 7 located in the following page presents the entire interpretive argument beginning from content design and leading to each of the three uses of the test.

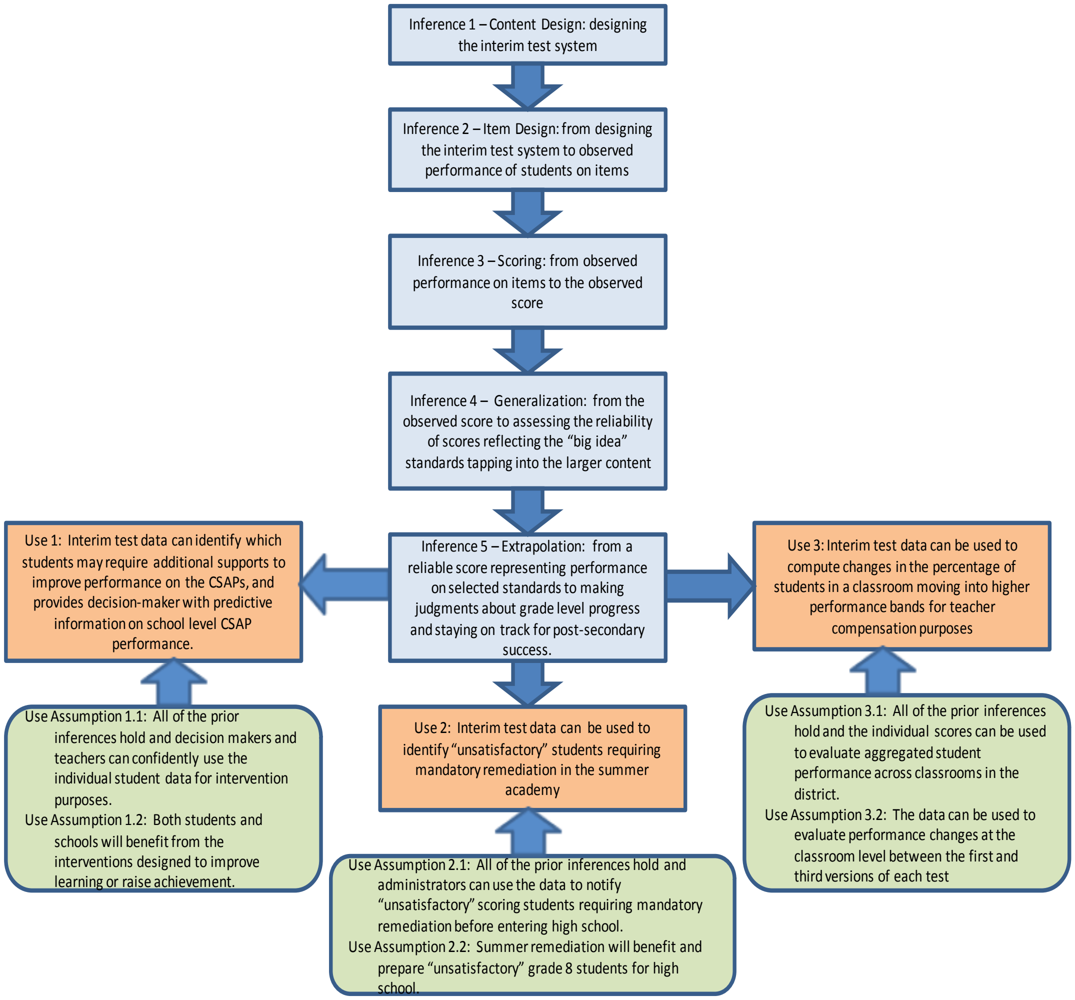


Figure 7. Interpretive argument for predictive uses, mandatory remediation and teacher merit pay

As seen in Figure 7, the assumptions under each Use inference depend entirely on the proficiency classification of students evaluated under both generalization and extrapolation. Use 1 presented in Figure 7 represents a common use of interim assessments in the country: the use of interim assesments as predictive tools for evaluating how well students will perform on the high-stakes end of year state tests (Perie et al., 2009; Yeh, 2006; Herman and Baker, 2005). Two primary ideas motivating the use of tests as a predictive instrument are that the tests identify which students may require additional tutoring or assistance prior to the high stakes testing window, and that the tests give an overall performance picture of how different classroooms or schools will most likely perform on the the high-stakes test. Regarding the latter point, the overall performance picture gives administrators at the school or district level a sense of the extent to which resources need to be mobilized to help improve these low performing classrooms and schools. At either the school or district level, the aggregated performance picture from interim assessments results would show administrators where either classrooms or schools are falling short of reaching district, state, and federal acountability standards. As noted in Chapter 1, in the case of this school district and many other urban school disricts populated with higher numbers of low achieving schools, this predictive perspective is valued by administrators facing public pressure to increase test scores and to produce better achievement results on the high stakes test each year. Evidence supporting the uses of the test under Use 1 comes from the evidence gathered previously to demonstrate that all prior inferences and assumptions evaluated appear to hold. In particular, the evidence should show that there is a strong relationship or association found beween the interim tests and the external criterion (CSAP) assessed under the generalization inference. Other evidence supporting the assumptions under Use 1 would also come from weighing the consequential aspects of using this interim assessment to fulfil the interventions or other actions instituted in response to the predictive information obtained from the test.

Use 2 in Figure 7, reflects the decision to use the test for identifying students for mandatory remediation . This use is closely related to Use 1, in that the assumption is made that the data from the interim assessments can be used with a high degree of confidence to identify and allocate resources to students judged by the test as "at-risk" of academically failing. The action distinguishing this inference from Use 1, is that the data were used to identify unsatisfactory grade 8 students for mandatory attendance in a summer academy in 2007 prior to entering high school. The idea for holding the mandatory summer academy was to ensure that the lowest performing grade 8 students gained the necessary skills needed to begin high school successfully. For students, this particular use raises the stakes for them since identification as "unsatisfactory" based solely on their performance on either the math or reading interim tests during the first year of implementation meant having to forego a part of the summer break before the start of the school year. Evidence to support the use of the test for mandatory remediation also comes from the evidence gathered in prior studies checking the earlier inferences and assumptions in the interpretive argument. Within the context of this use, the tests should identify unsatisfactory students with reasonable precision under the generalization inference. Further, students considered to be "unsatisfactory" on this test should also comprise of largely students who are deemed "unsatisfactory" by an external criterion evaluating the same target domain (CSAP). In addition, data and information attesting to the benefits of attending mandatory remediation provide supportive evidence for this inference.

Use 3 also assumes that the interpretive argument from content design to extrapolation holds in order to use the interim tests to evaluate teacher effectiveness based on student growth. In constrast to the first two uses, the data used to evaluate Use 3 are based on the aggregated performance at the classroom level to determine whether teachers should receive a merit pay bonus for improving student learning between the two time points assessed.

As discussed in the previous section, the evaluation of the interpretive argument in Figure 6 and Figure 7 entails designing studies to evaluate the assumptions noted under each inference. Due to the fact that all evaluations are bounded by time and resource constraints, all aspects of the interpretive argument cannot be evaluated or thoroughly examined. In this dissertation study, the following areas of the interpretive argument were not evaluated or sufficiently analyzed: all of the assumptions under content design, the Use Assumptions 1.2 and 2.2, and extrapolation. In reference to content design, acknowledging how well the content of the test reflects the target domain in the interpretive argument is important for establishing the connection between the foundation or design of the test system and all other areas of the test system examined. However, in order to properly evaluate the assumptions supporting content design, a different set of research questions would need to be developed that are outside the scope of this study and expertise from content experts would be required to assess the alignment between each interim test, the target domain and the corresponding CSAP test. Examples of research questions to evaluate content design would include: To what extent are higher order skills defined by the blueprints? How well does content reflected in each test measure the target domain?

Use Assumptions 1.2 and 2.2 are critical to acknowledge in the interpretive argument since these arguments emphasize the importance of weighing the consequential aspects associated with each use.

Similar to the challenge of assessing content design, Use Assumptions 1.2 and 2.2 (see Figure 2) require developing an entirely new set of research questions to evaluate the extent to which those uses have resulted in positive outcomes for the individuals or classroom and school units (for Use 1.2) being impacted. For Use 1.2, separate studies could be conducted to understand the nature of interventions and reforms taking place and the extent to which those reforms and interventions have benefitted those being directly impacted by those actions. In the case of Use 2.2, separate studies could be conducted to evaluate whether students who attended mandatory remediation benefitted from the intervention. For example, the experiences of students attending mandatory remediation and their subsequent academic experience in high school may be documented and analyzed to determine whether they believed remediation prepared them for high school. Another possible study checking Use 2.2 could entail assessing how many of the students who attended mandatory remediation successfully completed grade 9. Due to the breadth of studies and different types of data required to check these two assumptions, this study does not evaluate these assumptions.

In regards to the extrapolation inference, as discussed in the earlier section, the limited set of items represented by the MC and CR items on each interim test and the one context (formal testing situation) used to make judgments about students on the target domain makes it difficult to provide a convincing argument that the interim test adequately taps into the target domain. Although evidence is gathered in this study to determine the extent to which the interim test provides similar information about student performance as another test system, the CSAP, tapping into the same target domain and how many items appear to tap into the key grade level standards and benchmarks, a thorough evaluation would also include comparing the performance of students on the interim tests to student performance in other types of contexts outside of a formal testing situation.

For all other inferences and assumptions in Figure 6 and Figure 7, the validity argument draws largely on technical approaches recommended by the AERA, APA, and NCME *Standards* to find evidence to support these three uses of the test. With the exception of Use 3, the methods used to evaluate the interpretive argument are specified in the next chapter and accompanied by findings from appraising the argument. The methods and findings pertaining to Use 3 are located in a separate chapter since the evaluation of 6c moves away from individual students to assessing outcomes relative to data aggregated to the classroom level.

In contrast to the first three uses of the test presented in Figure 7, the interpretive argument presented in Figure 8 located in the following page draws on a different set of inferences and assumptions for evaluating Use 4. This interpretive argument outlines the assumptions tested to evaluate the claim made by DPS assessment and curriculum staff that valuable diagnostic information provided by data from the interim tests would lead to the improvement of instructional practices to meet the needs of students – particularly those students falling below proficient.

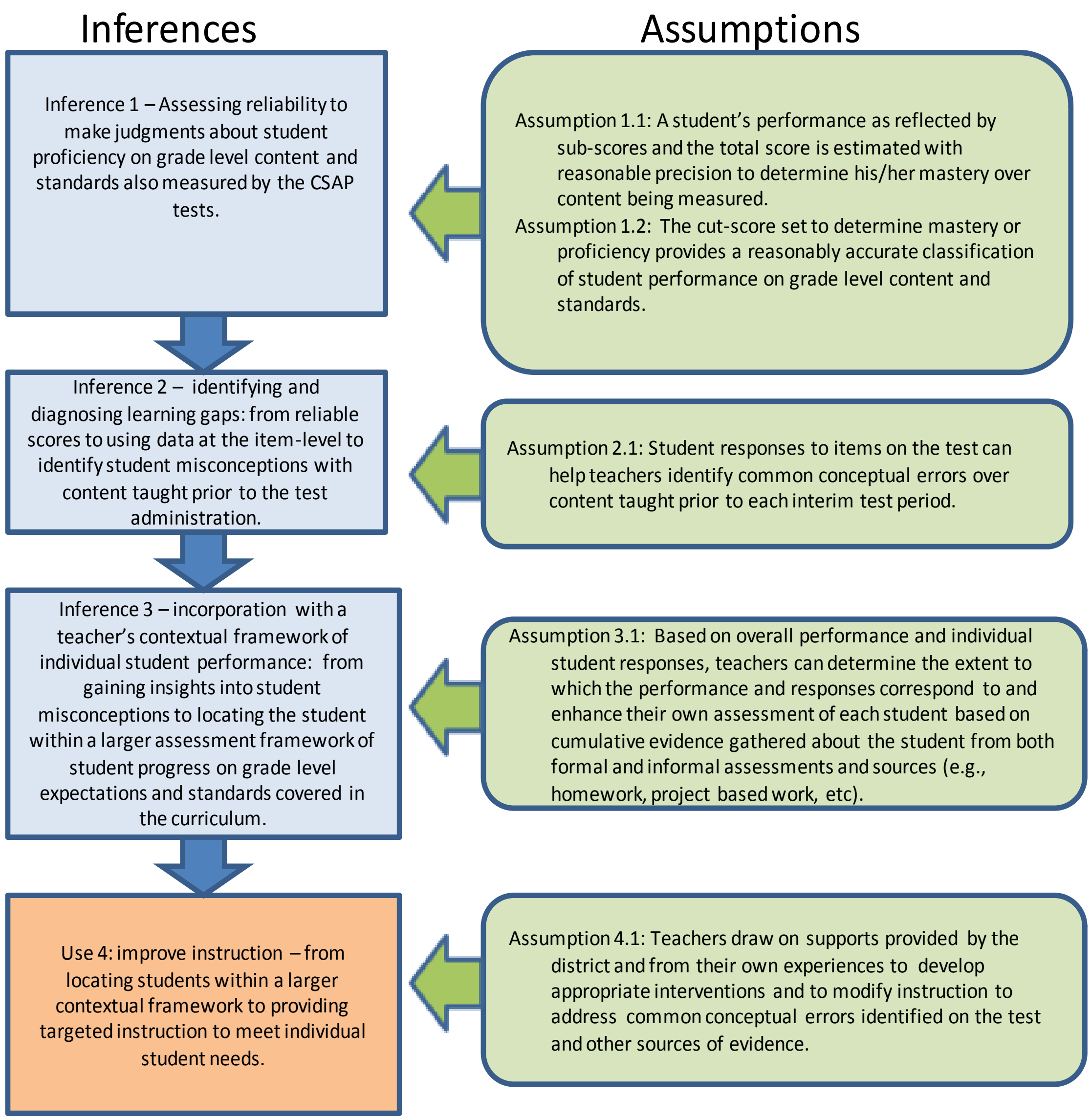


Figure 8. Interpretive argument for instructional use of the interim tests

The interpretive argument in Figure 8 follows the logic and assumptions of how developers of interim assessments envision teachers using the tests to first understand their students, and then to make the appropriate changes and adjustments in their instruction to meet student needs. In Figure 8, the first inference assumes that the diagnostic information provided by the sub-scores is reliable and that the performance cuts established represent a reasonable classification of students

The second inference in Figure 8 builds on the information gained from the overall proficiency and sub-score information gained about students in the first inference. That is, district assessment and curriculum staff expect that based on reviewing how students performed overall and on each of the larger standards represented on the test, teachers can then drill down to the items to gain more fine-grained diagnostic information about students. The inference about using item-level information to obtain better data about student misconceptions with material previously taught suggests that the test is comprised of items that can provide conceptual insights about student learning to teachers. In other words, teachers should be able to evaluate what aspects of their teaching needs to be adjusted to better address the root cause of student misconceptions over content. Ideally, these conceptual insights should be gleaned from the CR items requiring students to show their work.

After evaluating individual student responses to items, teachers can then align and check this information against their own contextual framework built largely from past and ongoing interactions with the student and from other bodies of evidence about what students know (e.g., homework and work completed during contact hours). The connections between how students perform on the test overall and sub-scores captured in the first inference of the interpretive argument, the individual items providing conceptual insights about student learning in the second inference, and a teacher's own contextual framework about each student, serves as the third inference of the interpretive argument. That is, this third inference reflects assumptions about the incorporation of interim assessment data into a teacher's contextual framework about each student. This incorporation is brought about by what Halverson (2010) refers to as the actuation stage, or the space made by teachers to reflect on new assessment data relative to all other evidence known about students. According to Halverson (2010), assessments from either formative or interim assessments can help enrich this contextual framework and provide teacher with valuable data points for helping them locate areas where instruction needs to be adjusted or modified to improve learning outcomes.

Following the third inference in the interpretive argument shown in Figure 8, the fourth inference assumes that teachers can then act on the information gained by comparing data from the interim assessments and other sources to make instructional improvements or adjustments as a means for improving student learning. According to the district, after reviewing data from the interim assessments and checking the information against their own contextual framework about each student, teachers should be able to make instructional modifications and adjustments as needed. Although test companies and institutions developing interim assessments commonly assert that the interim assessment data can be used to adjust instructional practices, studies by Heritage, Kim, Vendlinski and Herman (2009) and Blanc, Christman, Hugh, Mitchell and Travers (2010) found that when presented with data from these interim assessments, not all teachers know how to use the data. That is, some teachers did not routinely use data to alter instructional practices or did not know how to incorporate these data. Heritage et al. and Blanc et al. conclude that these findings suggest that training is critical for ensuring that teachers are given the tools for understanding how to use data from these assessments. Considering that the case study district is populated with many novice teachers who may not have the experience using data to drive instructional practices, the evaluation of this inference relied on gaining feedback from experienced teachers identified by central staff who have demonstrated expertise using data to inform their instructional practices.

The validity argument evaluating the interpretive argument in Figure 8 relies largely on evidence from interview data with veteran teachers to determine whether the assumptions supporting each link in the argument hold true. Although the interviews with teachers were structured to gather supportive evidence for each of the assumptions in Figure 3, the amount of time dedicated to each interview limited the extent to which each claim or assumption could be thoroughly investigated. For example, under inference 2 in Figure 8, evidence gathered to support this assumption came directly from teachers who were asked to provide specific examples of diagnostic uses of the test and to provide examples of how they identified student misconceptions with content taught previously. Although this approach can shed light on the conceptual insights gained from these tests, teachers were not asked during the interview to provide feedback on the richness of specific or individual items. Since the validity study for Figure 8 appraises an interpretive argument which differs substantively from the first three uses of the test, the findings from evaluating the assumptions in the figure are located in a separate chapter in this dissertation.

Limitations to the Validation Study

A key point emphasized by Kane in "Validation" and made earlier by Cronbach (1988) is that the criteria (clear, coherent and plausible) for appraising the interpretive arguments and the set of studies undertaken to evaluate test use are always open to further, or future challenges. The act of examining and testing out the arguments parallels the process of examining lines of scientific theory and similar to any scientific investigation, the findings are always open and subject to scrutiny and challenge. Further, because the scope of any evaluation is limited by cost and resource considerations, the limited scope in-itself implies that a validity study using an evaluation arguments approach cannot provide definitive or conclusive evidence but can provide compelling evidence suggesting whether proposed or applied uses of a test are justified. To help evaluators select which empirical studies should be pursued in a validity study, Kane proposes guiding the selection process by using four criteria outlined earlier by Cronbach (1989):

1. Prior uncertainty: is the issue genuinely in doubt?
2. Information yield: how much uncertainty will remain at the end of a feasible study?
3. Cost: how expensive is the investigation in time and dollars?
4. Leverage: how critical is the information for achieving consensus in the relevant audience? (pg. 26).

Again, although the above criteria may guide the process of designing a validity study, the necessary boundaries set by undertaking any evaluation study implies that definitive or conclusive information cannot be secured due to these constraints. Within the context of this study, the assumptions evaluated consisted largely of assumptions more easily checked with the data made available by the district. As noted earlier, the large scope and breadth required to check a few of the assumptions acknowledged in the interpretive argument resulted in omitting those assumptions from the validity argument. Before the findings from evaluating the interpretive argument are discussed, the next section of this chapter reviews five studies evaluating the different uses of interim assessments. These studies were selected since they present findings that are relevant to this study and to the larger question of whether these interim assessments appear to benefit students, teachers, and other stakeholders.

Interim Assessment Studies

As noted in Chapter 1, Shepard (2009) and more recently other educational researchers (Bulkley et al., 2010) note that very little research has been done to evaluate the uses of interim assessments. In this section, five studies of interim assessments that are instructive to the findings in this study are reviewed. These studies are: Vendilinksi's et al (2007) evaluation of items used on a set of interim science assessments developed by teachers and a group of educational researchers; Olah, Lawrence and Riggan (2010) and Clune and White's (2008) studies examining the instructional uses of interim assessments in case study urban districts; and, Nunnery, Ross and Goldfeder (2003) and Henderson, Petrosino, Guckenburg and Hamilton's (2007, 2008) studies evaluating whether higher achievement resulted in schools that implemented the same interim assessment system.

The first study reviewed in this section evaluates the quality of items used in four science interim assessments. This study represented the first phase of a three-part validity study of these interim assessments. In this study, the authors utilize the standards of the high-stakes end of year test administered in California to evaluate the qualities of the interim tests developed. The interpretive argument for evaluating the three out of four different uses in this dissertation follows Vendilinski et al.'s lead by also using the standards used by the state assessment to appraise the quality of items used in the DPS interim assessments.

The next two studies reviewed in this section represent studies evaluating instructional uses of the interim tests. These studies analyze interview and survey data to determine the extent to which the case study districts were using interim assessments to improve instruction. Similar to the test development process used by this school district, teachers in both case study districts were recruited to develop the interim assessments. In addition, similar to this case study district, the assessments used in those districts were rapidly developed and not piloted prior to administering the tests to students. That is, the assessments were produced within a few weeks and administered shortly to students after being published by the test developer. Considering the very similar processes in which these tests were developed in those studies and for this case study district, the findings from these two studies provide a good basis of comparison for the findings from exploring the instructional uses of the test in this study.

The last two studies reviewed in this section, Nunnery et al. (2003) and Henderson et al. (2007, 2008) evaluate the claim that the use of interim assessments can increase student achievement as assessed by the state standardized tests. Although these two studies examine a claim not evaluated in this dissertation, their findings contribute to the discussion in this study about the larger debate of whether investments in interim assessments appear to benefit students, teachers and school districts. These studies have been cited by other studies and district technical reports seeking to present evidence on the value of investing in interim assessments (for example, see Bulkley et al., 2010, Miami-Dade County Technical Brief, 2008; Yeh, 2006).

Evaluating the Technical Quality of Interim Assessments

The first study reviewed in this section was conducted by researchers of the Center for Research in Evaluation, Standards and Student Testing (CRESST) who evaluated whether a highly collaborative test development process could result in the creation of science assessments that the technical standards used by the high stakes end of year assessment. According to Vendilinski et al. (2007), this project was undertaken to develop the capacity of teachers to create high quality assessments needed to monitor student learning particularly since the high-stakes accountability environment requires better evaluation of student progress (pg. 2). This study represented the first of a phase of studies seeking to validate four science assessments for grades 4 and 5 developed by CRESST researchers and teachers. This first phase reviewed here focused exclusively on evaluating the quality of the items developed for the interim tests to determine the extent to which these tests could provide reliable information about students to district stakeholders. According to the authors, the second phase will evaluate whether these science assessments can predict future performance on the corresponding subject and grade tests on the high-stakes assessment system. The final and third phase of their study will assess whether the tests could be used to drive instruction in the classroom. To date, findings from the next two phases of the study have not yet been released.

For this study, a total of eighteen teachers and administrators[8] from five California school districts were selected to help create the science assessments with three CRESST researchers and five professional development leaders. Before the test development process began, all 18 of the teachers and administrators received three years of professional development to help deepen their understanding of assessment and content[9]. The twenty-six participants were divided into three groups assigned to initially develop tests for each of the three science assessments. Each panel was tasked with developing constructed response and multiple choice items for each assessment over a two and a half day period. The items were then piloted to selected grades 4 and 5 classrooms within five school districts in California. Data from the pilot were evaluated to determine whether the items met the same quality standards used to evaluate the items on the high-stakes standardized state test in California.

To evaluate the technical quality of the tests, Vendilinksi et al. used item point-biserials and Cronbach’s alpha to assess the reliability of each instrument. The Rasch model[10] was also used to evaluate the difficulty of the MC items. Items with point-biserials lower than the .3 standard used by the high-stakes assessment were flagged as requiring further review. The authors also used the .8 reliability standard used by the high-stakes assessment to evaluate the reliability of each assessment.

[8] The study does not inform readers on how many individuals were teachers out of the total number of 18 participants.

[9] The study does not provide information on how often these PD sessions took place over three year period. According to the paper, “Year 1 concentrated on large-scale assessment issues and involved teachers in the design, administration, and scoring of a multi-method assessment system designed for state use on the development and scoring of a science assessment called PASS...In Years 2 and 3, teachers applied the assessment framework described earlier to their classroom assessment practices…” (pg. 6).

[10]No rationale is given by the authors for using the Rasch model as opposed to other IRT models.

In general, the authors found that the collaborative structure of the test development process appeared to generate items meeting the standards used by the state standardized test. Each test assessed exceeded the reliability threshold of .8. At the item level, only five out of forty-one MC items for each test evaluated were deemed as exhibiting unacceptably high item-difficulty values and almost all items had point-biserials that met the acceptable range. Based on this finding, Vendilinski, et. al. contended that using collaborative processes and well defined guidelines and frameworks can lead to the creation of assessments that meet one of many criteria stipulated by Herman and Baker (2005) for validating interim assessments: the need to ensure that the items produce reliable estimates of student performance.

Vendilinski et al.'s study illuminates areas that contrast sharply from the test development process undertaken to develop the interim assessments developed for the district in this study. Unlike the rapid implementation of the interim assessment system in DPS, the teachers and administrators in Vendilinksi et al.'s study received training over a three year period to gain expertise in assessment development. According to the authors, "through a process of analyzing the results and determining implications for: (a) ongoing teaching and learning, and (b) refinement of the assessments and instructional plans for the next the unit was taught" (pg. 6). Vendilinksi et al. largely attribute the creation of quality items to the substantial time and resources devoted to professional development. Further, the researchers believe that the participating teachers have the added benefit of knowing how to analyze and use the assessment data based on the training received. The training in essence accomplished two objectives. The first objective was to provide item panel members with the training and guidance to ensure that the items reflected rich content and met specified technical standards. The second objective was to provide members with training to analyze and use data results in anticipation that these data may help improve achievement and instruction.

Another key point of contrast between the test development process used by DPS and the process used by the CRESST researchers is that a pilot phase took place to learn more about the quality of the items and to flag items that needed to be re-evaluated and possibly replaced. As noted in Chapter 1, the process of developing interim assessments varies between institutions and vendors developing these tests.

In the case of DPS and for the districts in the case studies reviewed next, all of these districts opted to administer these tests without piloting the items. The time taken by the authors to pilot the assessments allowed them to ensure that the technical quality of the tests met specific standards prior to expending the time and resources to administer the tests district-wide.

Evaluating Instructional Uses

The next two studies reviewed focused on whether teachers could improve instruction using interim assessments in case study districts. In both studies, the authors do not evaluate the content or the technical quality of the interim assessments, and focus the scope of their investigations to evaluating how teachers use the data from interim assessments to improve instructional practices. Both studies draw on predominantly qualitative approaches to evaluate the instructional uses of the interim test. Considering that the interim tests used in those case study districts closely resemble the interim assessments used in this case study district, the findings from these two studies provide a good basis of comparison for understanding the findings pertaining to instructional use in this dissertation study.

Olah et al. interviewed teachers to determine whether the interim assessments administered on a quarterly basis were changing and improving instructional practices in five elementary schools within one of the largest urban school districts in the country. The five schools were chosen on the basis that they all met adequate yearly progress in the 2004-2005 school year, represented a range of achievement on the end of year high-stakes math test, and shared "typical"[11] student demographic characteristics as the school district (pg. 228). Olah et al. limited their interviews to the grade levels (3 and 5) where students were assessed by the state assessment program. The interviews were conducted three times during the 2007-2008 school year, and took place in the Fall, Winter and Spring after the interim assessment reports were released to all schools. A total of 25 teachers across grades 3 and 4 classrooms participated in the interviews.

[11] Schools were selected on the basis that they shared similar percentages of free and reduced lunch eligible students, English language learners, students of color, and students with disabilities as the entire district.

Based on analyses of the interview data, Olah et al. state, "what is striking in our study of teachers' use of assessments is just that – teachers' use. As we have stated elsewhere, and it bears repeating here, teachers are using these assessments" (pg. 244). That is, Olah et al. found that teachers were reviewing the data from the score reports, looking at the standards where students were not performing well, and drilling down to individual items to learn what kind of mistakes were being made by students. However, Olah et al. found that when they asked teachers to expand on the quality of information obtained from the scoring reports and specific items, the data pointed to more superficial uses of the test to drive instructional decisions. For example, teachers often used the data largely to illuminate procedural mistakes made by students, and teachers stated that they modified their instructional practices to address these common procedural errors. Olah et al. found that none of the data points collected revealed that the interim tests were helping teachers locate where or why students may be struggling with specific concepts. Although the authors found that teacher uses of the interim assessments did not appear to conform to the diagnostic processes and formative practices that the interim assessments claim to foster and that district administrators envisioned, the authors remain optimistic about the benefits of the assessments since, "the fact remains that [teachers] are consulting, analyzing and acting on interim assessment results" (pg. 244).

In contrast to the Olah et al.'s findings, Clune and White's study (2008) found that teachers using the quarterly interim assessments in the Providence Public School District (PPSD) experienced difficulty using the assessments at the level described in Olah et al.'s study. Similar to the interim assessments used by the case study district in Olah et al.'s study, the assessments in PPSD comprised of quarterly assessments developed by teachers under the guidance of a testing company. Similar to the experience of DPS, the interim assessments in PPSD were used for two years (2004-2005 and 2005-2006 school years) before the district opted to no longer use the services of the test company who facilitated the development of these tests.

In this study, Clune and White evaluated whether the interim assessments were fulfilling three goals articulated by district administrators. These goals were to: 1) complete the alignment between the scope and sequence and grade-level expectations with the curriculum and assessments; 2) provide practice and preparation for the state assessment; and 3) provide data for teachers on the instructional needs of students (pg. 7). Unlike Olah et al.'s study and this dissertation study, the use of the interim assessments in PPSD to fulfill the instructional need objective was not framed explicitly by district staff in terms of getting diagnostic information about students. In their study, the authors were primarily interested in learning whether teachers were reviewing the data and modifying instruction without inquiring about the diagnostic quality of the assessments for improving instruction in the classroom.

The study design consisted of two rounds of gathering data through focus groups, interviews, and surveys. The first round was conducted during the first year of implementation (2005-2006) and the second round was conducted during the second year of implementation (2006-2007). For the first round, the researchers conducted two focus groups per level and at different school sites (a total of six focus groups) and ten interviews with district central staff members. For the second round, the researchers conducted 10 focus groups at the elementary level and three each at the middle and high school level (a total of 16 focus groups), and eight interviews with district central staff. The focus groups at school sites typically included a staff administrator (the principal, or assistant principal) and two to four teacher participants. According to Clune and White, the school sites were selected by both district staff and the Institute of Learning who wanted to ensure that the researchers visited schools with varying levels of buy-in for the interim assessments. The surveys were administered district-wide to ensure that perspectives from all schools were included in the study.

According to the authors, the interview, focus groups, and survey data provided evidence that all three objectives were being fulfilled to some extent but with large variability found in the frequency in which teachers were reviewing and using the data across levels. Data from elementary school sites showed more positive support for using the interim assessments and the belief that the interim assessments were useful measures of content taught in class, for predicting performance on the state test[12]

[12] The authors acknowledge that either authors or district personnel did not test opinions noted about the predictive characteristic of the interim assessments.

and for helping teachers identify weaknesses or gaps in learning. In contrast to the elementary level, the findings for the middle and high schools revealed that many teachers did not believe the assessments aligned with the curriculum and spent little time reviewing the results from the data. At all levels, however, teachers agreed that the two week lag separating the administration of the test and receipt of score reports made it difficult for teachers to re-teach content reflected on the assessment since teachers were already moving into new topics.

Based on the mixed findings from their study, the authors state, "…it was unclear whether the interim assessments functioned as a system of classroom assessment capable of producing the major gains in achievement attributed to formative classroom assessment" (pg. 14). That is, the authors could not obtain a clear picture or consensus of whether district staff and teachers believed that student performance was increasing as a result of using the interim tests. However, Clune and White state that they did not design their studies to evaluate whether student achievement was increasing and that a different study design and new research questions would need to be developed to test out this line of inquiry. The next two studies reviewed provide examples of studies designed to evaluate whether the use of interim assessments had an impact on student achievement as measured by the high stakes assessments.

Evaluating the Achievement Impact of Interim Assessments

The studies by Nunnery et al. (2003) and Henderson et al. (2007, 2008) test a common claim made by test developers that interim assessments can contribute to increased student achievement. These studies have been cited by other authors seeking to provide evidence supporting the use of interim assessment systems. Other commonly cited studies that also evaluate the relationship between assessments and achievement, such as the most frequently referenced Black and William (1998) review of formative assessment studies, do not evaluate assessment systems producing standardized results across classrooms, schools and districts as the basis for their studies. Setting aside studies evaluating assessments that do not conform to the characteristics of interim assessments is important to ensure that achievement gains typically attributed to classroom based or generated assessments are not conflated with the achievement gains made using interim assessments with a standardized focus. One key difference between the Nunnery et al. and Henderson et al. studies is that the former study clearly establishes that they are interested in detecting the impact of a specific type of interim assessment program which required all schools within a school district to change the curricular focus in literacy and math. Henderson et al. on the other hand, focused their study on the question of whether investing in interim assessments leads to higher achievement gains.

In Nunnery et al.'s study, the authors compared the performance of nine elementary and two middle schools located in one district on the reading and math Texas Assessment of Academic Skills (TAAS) relative to the same number of comparison schools located in other districts during the first three years of implementation. The eleven schools situated in the same district represented schools that implemented the same set of math and reading interim assessments from one provider. Program schools were matched with comparison schools based entirely on student characteristics such as the percentage of free and reduced lunch eligible students, limited English proficient learners, and students belonging to each of the five major racial/ethnic groups at each school site. The authors first conducted a set of cross-sectional studies that focused on year to year performance comparisons between students in program and comparison schools. The cross-sectional analysis consisted of three parts: comparing the mean achievement scores by subject area and at each time point between each program and comparison schools by grade, comparing the median effect size estimates by grade and year across all program schools relative to all comparison schools, and comparing the percentage of students performing at or above grade level by grade each year. These analyses were used to provide a general profile or overview of performance by all students in each grade and at each of the three years reviewed in this study.

The second set of studies consisted of a repeated-measures analyses to compare achievement on the reading and math TAAS between two cohorts of students in program and comparison schools from the first year of implementation (1989-99 for reading, 1999-00 for math) to the third year of implementation. The first cohort represented students moving from grades 3 through 5 and the second cohort represented students who moved from grades 5 through 8. Unlike the cross-sectional studies, the repeated-measures analyses took into account a student's prior performance and socioeconomic status. Supplementary analyses conducted by the authors included a comparison of subgroup performance (free and reduced lunch eligible students, limited English proficient students, and students below proficient) and performance relative to level of reform implementation at each program. Since the authors state that findings from the supplementary analyses were conducted using very small sample sizes and that the findings from these analyses need to be interpreted with great caution, this review only discusses the main findings presented in this study.

The findings presented from running both longitudinal and cross-section analyses are mixed. For the cross-sectional analyses, despite the fact that a few grades showed statistically significant differences between the two groups of students, the effect size difference was virtually "near zero" for each grade. Further, the percentage of students considered to be at or above grade level by year three of implementation were either virtually identical or did not differ by more than two percentage points in each grade (grades 3 through 8) and in each year assessed. For the repeated-measures analyses, the findings revealed that, the elementary cohort of students enrolled in program schools for three consecutive years outperformed their peers in both reading and math with notable effect size differences of .22 for reading and .2 for math. The middle school cohort in program schools performed at a very similar rate to their comparison peers in reading, but slightly outperformed their peers in math. For math, the effect size difference between the program school cohort relative to the comparison school cohort was .17. For all the repeated-measures analysis, program type only accounted for no more than 1.8 percent of the variability in scores found across cohorts being compared.

Although the authors conclude that, "taken as a whole, the present results are clearly suggestive of its benefits for student achievement, and if consistently replicated in future studies would strongly imply proven effectiveness", the practical and substantive significance of the findings do not provide a clear picture of the benefits of this particular interim assessment. In general, the findings based on the cross-sectional analyses show similar performance outcomes between the two groups, and the repeated-measures analyses appear to suggest that the interim assessments were more effective in raising elementary school achievement at program schools, but were not as effective in increasing performance in the middle schools. The findings as presented in this report do not provide clear-cut evidence on whether districts should invest in interim assessments as an effective strategy for raising the achievement of all students.

In contrast to Nunnery et al.'s mixed findings, the two-year study by Henderson et al. did not detect statistically significant differences in achievement gains made by a group of 22 middle schools that received a grant to invest in the same interim assessment system and 44 middle schools that did not implement the same set of interim assessments. The authors selected comparison schools on the basis that they shared similar student characteristics and almost "identical" achievement as the schools with interim assessments prior to the first year of implementation. For the first year of the study, the authors wanted to learn whether there were any immediate achievement effects that could be detected in schools with the interim assessment system after the first year of implementation. The authors used five years of achievement data from the 2000-2001 school year to the 2005-2006 year to compare achievement at each time point and across the years for both groups. In year two of the study, the authors evaluated whether any achievement gains could be detected after using the interim assessment system for two years. The authors ran t-tests to compare average reading and scale scores between groups at each time point and conducted an interrupted time series study to compare growth trajectories for both groups over the six year time period in the first study, and over the seven year time period in the second study.

Henderson et al. found that evaluating the data from a status and longitudinal perspective yielded the same results: program schools and comparison schools shared similar performance at each time point and across years. When comparing status achievement, the program schools performed slightly higher than the comparison schools, but the authors found that this difference of less than one scale score point between groups was not statistically significant. The interrupted time series analysis found that both groups experienced an achievement increase that was statistically significant in the 2005-2006 year relative to all prior years. However, when comparing whether the one-year achievement gain differed significantly between groups, the authors found that these gains were virtually identical. The authors conducted the same study for the second year and found identical results after two years of program implementation. That is, program and comparison schools exhibited similar status achievement and growth after year two.

Limitations to this study that are also acknowledged by the authors include the small size of the program group relative to the comparison group, the limited years of data evaluated to make judgments about effects associated with the interim assessments, and the fact that the authors did not investigate whether the comparison schools may have been using other interim assessments. The last limitation briefly noted in the studies potentially invalidates the findings that were designed to evaluate whether the implementation of interim assessments can increase student achievement. That is, this research question cannot be answered if schools in the comparison group are using other types of interim assessment programs.

Although this dissertation study does not evaluate the specific claim that interim assessments can contribute to higher achievement outcomes on the high stakes test, all of the uses evaluated in this study and in all other studies of interim assessment have the larger objective of ultimately increasing student achievement as measured by the high-stakes test. As noted in Chapter 1, this school district, like many other urban school districts, faces tremendous pressure and public scrutiny to increase student achievement. The studies from Nunnery et al. and Henderson et al. studies do not provide clear answers to the policy question of whether investing in interim assessments serves as an effective strategy for raising achievement. Further, the findings from the other three studies reviewed in this section reveal that more evidence is needed to help districts understand the added benefits of these interim assessment programs. Vendilinski et al.'s study showed that interim assessment tests can be carefully constructed to develop highly reliable tests, but whether these tests of high technical quality are useful to stakeholders and can predict performance on the state tests remain as open questions. Olah et al., found that the interim assessments used in the case study district proved useful to teachers to gain a general indicator of overall achievement, however they found that teachers could not use the data from the interim assessments to learn about student misconceptions with content and gain insights into their instructional practice – which was a critical objective for implementing these assessments in the district. Clune and White also found that some teachers believed that the assessments provided valuable information about their students, but found that many teachers in the middle and high school grades did not review the interim assessment data and felt that the tests were not aligned with the curriculum. Taken together, the findings from these studies suggest that far more research needs to be conducted to evaluate the uses of interim assessments and determine the justification of investing in such test systems to fulfill district objectives.

Clune and White states, "interim assessments are not cheap…adopting a policy [of investing in interim assessments] because it is possible and might achieve any one of several good purposes is tempting and may explain the rapid growth of interim assessments". Yet, despite the well-acknowledged paucity of research studies showing the effectiveness of these assessments to meet common uses of these tests, this trend of districts investing in these systems not only continues to rise but the stakes associated with data from interim assessment systems have also increased over time. The studies in the next three consecutive chapters seek to contribute to this small research base by evaluating the extent to which the interim assessments in this case study district appear to support uses that are also common to many school districts across the country.

CHAPTER 4 – Evaluating the Interpretive Argument

The focus in this chapter and the one that follows is on evaluating the interpretive argument illustrated earlier in Figure 7. The evaluation of the interpretive argument provides supportive evidence for the validation argument or the outcome after the inferences in Figure 7 shared by all three of these uses have been evaluated. Figure 7 introduced in the previous chapter, addresses the use of the interim tests for meeting predictive purposes, for mandatory remediation and for compensating teachers under the merit pay system.

As discussed earlier, the set of analyses contained in this chapter begin with evaluating the second inference or item design through the generalization inference, and culminate with evaluating how well the evidence from all inferences meets the specified uses. The first inference and the extrapolation inference are not evaluated in this dissertation study. In addition, Use 3.2 in Figure 7 is evaluated separately in the next chapter since that particular assumption rests on decisions being made at the aggregated classroom level and only applies to the third use of the test. As indicated earlier, Figure 6 presents the set of assumptions supporting each of the common inferences shared across all three uses. Findings presented here follow the same sequence mapped out in Figure 6.

Methods for Appraising the Interpretive Argument

The analyses conducted in this chapter rely primarily on using classical test statistics and Item Response Theory (IRT) analyses for one test: version 2 of the grade 8 math, version 2, test[13] administered at the end of the first semester in the 2006-07 school year. The grade 8 math test was chosen to empirically test the assumptions supporting the interpretive argument since this particular test addresses two of three uses specified in Figure 9. An IRT approach, specifically the partial-credit model (PCM; Masters, 1982) was used to appraise different areas of the interpretive argument since this approach

[13] From this point onwards, this test is referred to throughout this chapter as the “grade 8 math” test.

illuminates whether students of varying proficiency are measured with adequate precision, particularly in relationship to the established proficiency cut-scores.

Classical test statistics (item difficulty and item-total correlations) are conventionally used to report information about the properties of a given test. Although these statistics are reported in this chapter, the PCM was used to gain information not readily accessible under classical test theory (CTT). Under CTT, a respondent with a high observed score and a respondent with a low observed score are assessed with the same amount of error[14] by a test instrument. IRT, in contrast to CTT, assumes that a latent variable underlies both the items and respondents and uses the same scale to estimate the locations of both respondents and items along this variable. The measurement error or the standard error of measurement (SEM) in IRT varies with the associated estimates for all respondents being measured by the test[15]. The IRT feature of placing the estimated locations of respondents and items on the same scale allows for a direct comparison to be made between a respondent's proficiency level and the difficulty of a test's item. The degree to which the SEM varies for respondents depends on the location of items at different points along the scale (i.e, item's difficulty), and the number of items found at these different locations. As discussed earlier, since the impact from the uses evaluated from this study is largely determined by the student's score on this test, estimating the precision with which different types of students are being evaluated is important. The SEM is discussed in more detail later in this section and used to evaluate the item design and generalization inferences.

The PCM used to evaluate data from the grade 8 math test comes from the Rasch family of IRT models. In addition to estimating respondent and item locations on the same scale and allowing the SEM to vary, the PCM is also used because this model can accommodate tests with mixed item formats (i.e., both multiple-choice (MC) and constructed response (CR) items). When generating respondent estimates and item location estimates for MC items, the PCM takes on the same form as the Rasch model since

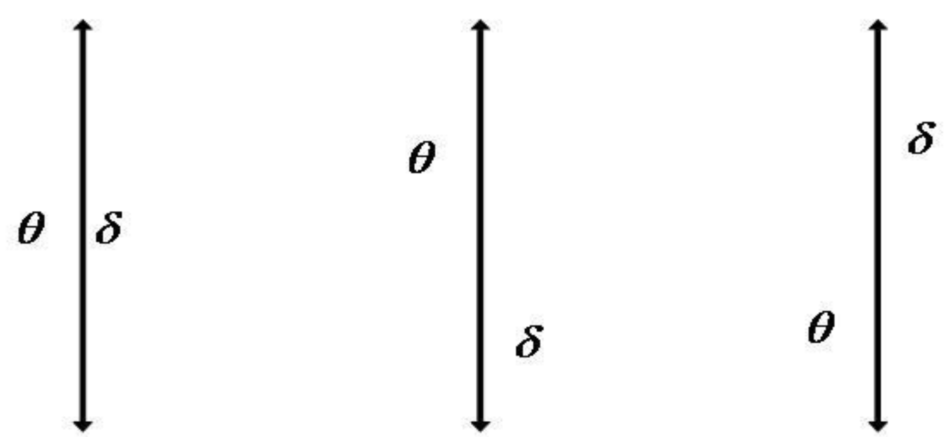
[14] See pages 34-35 for a formal definition of what is meant by the use of the term "error."

[15] The SEM can also vary for items, but the focus of the SEM discussed is on evaluating how well the items measure respondents.

there are only 2 response categories (a score of 0 or 1) to consider. For CR items with greater than 2 response categories, the Rasch model can be extended to the PCM. The PCM was modeled in this study using the BEAR Assessment Center's ConstructMap version 4.5.0. The following explanation of both models is based on Wilson's (2005) presentation of the Rasch and PCM models in *Constructing Measures: An Item Response Approach*.

The Rasch and PCM

The basic premise of the Rasch model is that the relationship between both respondents and MC items can be understood relative to the difficulty of the item. Under a Rasch model, both respondents and items are estimated using the same standardized logit scale. Figure 11 taken from *Constructing Measures: An Item Response Approach* (Wilson, 2005) illustrates the relationship between respondent and item difficulty expressed by the Rasch model.



(a) (b) (c)

Figure 9. Representation of three relationships between respondent location and the location of an item (from Wilson, 2005)

Figure 9 presents three different scenarios between a respondent and an item's locations. In Figure 9, the symbol, θ (theta), represents a respondent's estimated proficiency and the symbol, δ (delta), represents an item's estimated difficulty level. The thick line between the theta and delta symbols represents the latent variable measured by a test. Estimates of both item difficulty and respondent proficiency increase moving from bottom to top along this latent variable. Scenario (a) shows theta and delta at the same location. Scenario (b) shows theta at a higher location than delta, and scenario (c) shows theta located at a lower level than delta. Under the Rasch model, if $\theta = \delta$ then a respondent has approximately a 50% chance of answering the MC item correctly. If $\theta > \delta$ then a respondent has more than a 50% chance of answering the item correctly, and if $\theta < \delta$ then the respondent has less than a 50% chance of answering the item correctly. The equation underlying the Rasch model scenarios presented in Figure 9 is presented below:

$$\log\left(\frac{P(X_i = 1)}{P(X_i = 0)}\right) = \theta - \delta_i \qquad (4.1)$$

Equation (4.1) indicates that the log of the odds of a correct item response is equal to theta minus delta. Since the log-odds are also commonly referred to as logits, Equation (4.1) can be simplified to the following expression:

$$\text{logit } (1{:}0) = \theta - \delta_i \qquad (4.2)$$

In Equation (4.2), if $\theta - \delta_i$ is equal to 0 logits, this would mean that both theta and delta are of equal value and that this individual has a 1:1 odds ratio or a 50% probability of getting this item correct.

The three relationships depicted in Figure 9 along with all other items used on a test instrument, can be visually represented through a *Wright Map*. A Wright Map combines information from the Rasch model for all items on a test with the idea of an underlying latent variable being measured by the test. Figure 10 illustrates a Wright Map produced for one interim assessment in DPS measuring grade content standards and benchmarks in math.

```
=============================================================================

   raw      students                                item.step
-----------------------------------------------------------------------------
    |30                  XX|
3   |                      |
    |29              XXXXXX|MC5.1
    |                      |
    |28             XXXXXXX|
    |                      |
    |27             XXXXXXX|
2   |26            XXXXXXXX|
    |25            XXXXXXXX|CR18.1________________________advanced
    |24           XXXXXXXXX|
    |23           XXXXXXXXX|CR4.3 CR12.3 MC19.1
    |22            XXXXXXXX|CR4.1
    |21           XXXXXXXXX|CR4.4 CR6.1
1   |    XXXXXXXXXXXXXXXXXX|CR18.2
    |18          XXXXXXXXXX|CR4.2 MC7.1
    |17 ---------XXXXXXXXXX|MC11.1 CR12.2 MC23.1
    |16          XXXXXXXXXX|MC8.1 MC21.1
    |  XXXXXXXXXXXXXXXXXXXX|___________________________proficient
0   |13                   X|
    |12           XXXXXXXXX|
    |11           XXXXXXXXX|MC13.1 MC16.1 MC20.1____partially prof
    |10           XXXXXXXXX|CR6.2 MC15.1
    |9            XXXXXXXXX|MC2.1 MC3.1
    |8              XXXXXXX|CR12.1
-1  |7              XXXXXXX|MC22.1
    |6                 XXXX|MC1.1 MC17.1
    |5                 XXXX|MC9.1
    |4                  XXX|
    |                      |MC14.1
    |3                   XX|MC10.1
-2  |2                    X|
    |                      |
    |1                    X|
    |                      |
    |0                    X|
    |                      |
=============================================================================
```

Figure 10. Wright Map of grade 4, math, 2007-08 test

Each “x” located on the left-hand side of the dashed line dividing items from respondents represents a group of 32 students with the same estimated proficiency level. Each x represents a group of respondents located at the same proficiency. The items are noted by their number and response or step category on the right hand side of the dashed line separating the items from the respondent distribution. MC indicates that the item is a multiple-choice item and CR indicates that the item is a constructed-response item. For example, MC11.1 represents multiple-choice item 11 and a response score of 1 and item CR4.3 represents constructed response item 4 and a response score of 3. As noted earlier, both proficiency and item locations are expressed and estimated using the same scale, the logit scale, and reference to this scale is located on the far left hand side of the Wright Map. The logit scale typically ranges between -3 and 3. Higher theta values (e.g., above 1 logits) represent respondents who have demonstrated more mastery over the content being measured, and subsequently, more difficult items would also be located higher up on the logit scale. The raw score is also provided to the right of the logit scale as another point of reference. The Wright can be used to estimate the odds or the probability for a respondent to answer an item correctly by evaluating the distance between a given respondent and item on the map. As depicted earlier in Figure 9 and expressed in Equation (4.2), if both item and respondent are located at the same logit value, then the respondent possessing a particular level of proficiency has a 50 percent probability of selecting the correct response on the item representing a specific level of difficulty.

One advantage of the Rasch model just described is that one can always calculate the probability of a respondent getting an item correct by simply knowing the item's estimated difficulty and the respondent's estimated location. At the instrument level, the probability of a person's entire set of responses to an instrument (the response vector) that consists of only MC items can be calculated by first: assuming that each item independently contributes to this response vector, calculating the probability of correctly endorsing each item (convert the logit values into probabilities), and then taking the product of all item probabilities. For example, if a respondent has a response vector to three items of 0, 1, and 0. If the probability of responding 0 on the first item is equal to .5, the probability of responding 1 to item 2 is .8 and the probability of responding 0 to item 3 is .7, then under the conditional independence assumption, the probability for this response vector is equal to (.5)(.8)(.7) or .28. The assumption of an item contributing independently to a response vector is called the conditional independence assumption and would most likely not hold if a few items share the same stem or "stimulus" materials.

The PCM extends the Rasch model presented in Equation (4.2) to accommodate additional response categories. Under a PCM, the steps between the response categories can vary in distance such that a student needs to exhibit more proficiency or mastery over the content being measured in order to move from one step to the next, or that very little proficiency or mastery is required to move from one response category to the next. Recall that the Rasch model uses the formula below to determine the odds that a student would score a 1 on a given item:

$$\text{logit}\ (1{:}0) = \theta - \delta_i \tag{4.3}$$

To move from the dichotomous to the polytomous (more than 2 responses) context, Equation (4.3) can be modified for all additional step categories by substituting the odds of scoring a 1 to other pair of scores. For example, for an item with 4 possible scores (0, 1, 2, 3), the logit relationship expressed in Equation (4.3) or the dichotomous context of 0 and 1 would also apply to each pair of ordered scores such as 1 and 2 or 2 and 3 or 3 and 4. For the paired scores of 3 and 4, Equation (4.3) would be modified to the following equation:

$$\text{logit (4:3)} = \theta - \delta_{i4} \qquad (4.4)$$

The interpretation of Equation (4.4) is that the relative probability for a person who scored either a 3 or a 4 to score the higher response of 4 is a function of the difference between her location and the item difficulty for that "step parameter". Under the polytomous context, the parameter δ_{ik} or the "step parameter" represents the probability of moving from the step or response category from score k-1 to 1. The higher the value of a step parameter *(* δ_{ik} *)*, the more difficult the step is relative to lower steps within a given CR item.

To illustrate, the item characteristic curves (ICCs) for one CR item in an interim test is presented in Figure 11 and depicts the relationship between an estimated proficiency (θ) and where step categories intersect at δ_{ik} . The x-axis represents the varying levels of θ and the dotted lines represent the threshold or each δ_{ik} where a lower category intersects with the next higher response category. The category response curves here are ordered from left to right beginning with the first category p(0) or a score of 0 and ending with the last category p(4) or the highest score of 4.

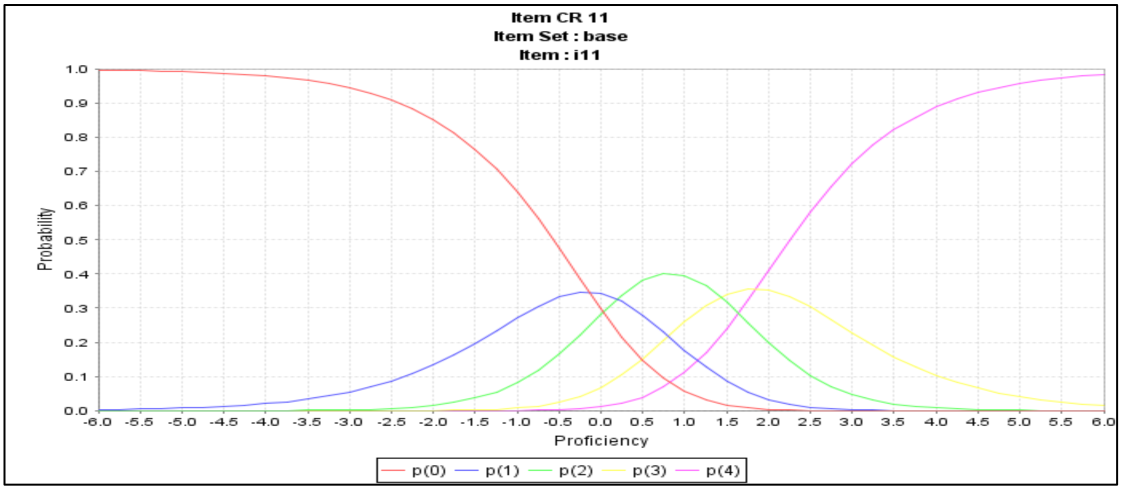


Figure 11. Item response curves for a CR item on a grade 4 interim test

A student estimated at -2.0 logits has a very high probability (approximately 85% probability) of scoring 0 on this item. A student located at 1.0 logits would most likely score a 2 since she is located well above the thresholds for scoring either a 0 or 1 and for scoring either a 1 or a 2. However, because she is located well below the threshold of having a equal probability of scoring either a 2 or a 3, she is less likely to score a 3. Figure 11 also shows that the distance required to move from a score of 1 to 2 to a score of 2 to 3 is considerably farther from other thresholds and this suggests that a student must demonstrate substantially more skill or proficiency to score higher than a 2 on this particular item. As noted earlier, in the case of MC items, the PCM takes on the form of the Rasch model represented in the simplified Equation (4.2) since there are only two responses (0 or a 1) or one threshold separating two possible responses.

In Figure 11, the step categories are ordered from lowest to highest along theta. However, this order can sometimes be reversed as seen with several of the CR items shown in the Wright Map presented in Figure 10. For example, Item 18 shows a reversal in response categories between the highest response for that item (CR.18.2) and the lower response category (CR18.1). As seen in the Wright Map, the one-point response category is located higher than the two point category and this would suggest that students at lower levels of proficiency have a higher probability of earning full points on this item and a much lower probability of earning 1 point on this item. Embretson and Reise (2000) point out that these reversals may suggest that there is one response category that respondents will seldom achieve regardless of proficiency level. In the case of this specific grade 4 test, this would indicate that students were more likely to score a 0 and 2 on CR item 18 but were less likely to receive a score of 1 on this particular item. Considering that CR items are supposed to provide more information about students, and that there are a limited number of items on each interim test, CR items that fail to differentiate varying levels of performance should undergo evaluation and revision. That is, if very few students land in a response category for a given CR item, this finding would defeat the purpose of investing in these considerably more expensive items to better differentiate the proficiency of students over content assessed by each test.

In order to obtain item parameters and respondent proficiency estimates, the estimates need to be anchored to either the distribution of items or respondents (Embretson & Reise, 2000). That is, each point within the scale is given meaning relative to the mean and standard deviation of the distribution of either items or respondents. In ConstructMap, the logit scale is anchored to the item distribution with the mean item difficulty set to 0 and the standard deviation set to 1. Anchoring the solution to items in this manner references the performance of students to the item distribution. Conversely, if the scale had been anchored to the respondent proficiency distribution, the difficulty of items would be referenced against the distribution of student proficiency. Therefore, if a student is located well above 0 logits, she would have a higher probability of getting most items correct on the assessment. If a student is located at exactly 0 logits, then that student's performance can be characterized as being at the average difficulty level of all items on the assessment. The next section specifies the methods used to assess the reliability and the standard error of measurement reported in this chapter under the item design inference and the generalization inference.

Evaluating Reliability and the Standard Error of Measurement

In this chapter, the internal consistency for each of the nine tests (see Table 10 on page 115) is evaluated using a common reliability statistic: Cronbach's alpha. Cronbach's alpha represents a test of internal consistency which evaluates how well a test appears to be measuring students reliably. Equation (4.5) presents the formula for Cronbach's alpha:

$$\alpha = \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N}\sigma_{y_i}^2}{\sigma_x^2}\right] \tag{4.5}$$

Where N is equal to the number of items, σ_x^2 is equal to the variance of the overall scores on the test, and $\sum \sigma^2 Y_i$ is equal to the sum of the variance on scores for each item (i). For the expression enclosed in brackets, if the denominator or the sum of the variance of the overall scores is large, then this would result in a smaller ratio. A smaller ratio, subtracted from 1 would subsequently yield a higher alpha level. A large variance found in overall scores means that there is more spread or variability in total scores earned by students and this characteristic is desirable since this means that the instrument is better differentiating between students of varying proficiency. Therefore, a test designed with homogenous items (all easy or all difficult) would most likely be low in reliability since the items would not be sensitive to the performance of different types of students.

The alpha scale typically ranges from a scale of 0 to 1, with 1 representing perfect reliability. The conventional standard within the educational testing context for assessing whether scores on an instrument are reliable is approximately .8. A test with a reliability found to be lower than this threshold suggests that the test may not be providing consistent measures of respondents.

In contrast to Cronbach's alpha, the test information function or the conditional standard error of measurement (SEM) reported in IRT is typically used to improve the instrument by evaluating the extent to which respondents located at varying levels of the proficiency distribution are being assessed with more precision. In this study, conditional SEMs are used to evaluate whether sufficient items are available to measure respondents at different proficiency locations under the item design and generalization inference. Unlike the conventional assumption under CTT where the same constant SEM applies to all respondents, in IRT the SEM varies depending upon the estimated location of a respondent. For illustrative purposes, Figure 12 shows a SEM curve for students on a grade four interim test.

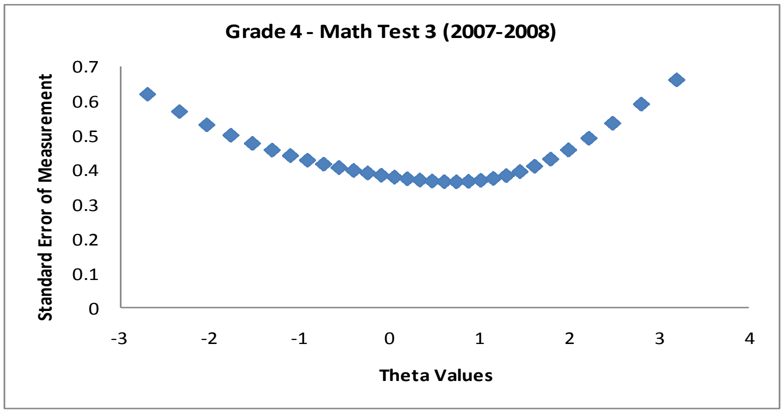


Figure 12. Standard error of measurement for one interim test

Note that the SEM typically resembles a u-shaped curve where students located at the extremes tend to be measured with more error than those located at the center of the respondent distribution where more items contribute to improving the estimates of those respondents. As shown in Figure 12, students located at the lower ends of each distribution have higher error estimates associated with their scores and students at approximately 0 logits have a smaller SEM estimated with their scores. The SEM is used to evaluate cut-points commonly found on most achievement tests to designate proficiency cuts. For example, for the same grade 4 math test depicted in the SEM plot, a student labeled in the upper region of "partially proficient" would be located at .05 logits with an associated raw score of 13. As indicated by Figure 12, this respondent's score is located 2 points below the proficient cut point set at a raw score of 15. In the logit metric of IRT the SEM associated with this person's estimate is .38. When a confidence interval is placed around this person's estimate, a SEM of .38 means that the respondent could be located anywhere between -.33 and .43 logits (plus or minus .38 from the respondent's score).

```
    raw       students                                       item.step
------------------------------------------------------------------------------
    |30                         |
3   |                           |
    |29                         |i5.1 i5
    |                           |
    |28                         |
    |                           |
    |27                         |
2   |26                         |
    |25                         |i18.1_________________________advanced
    |24                         |
    |23                         |i4.3 i12.3 i19.1 i18 i19
    |22                         |i4.1
    |21                         |i4.4 i6.1 i4
1   |                           |i18.2
    |18                         |i4.2 i7.1 i7
    |17                         |i11.1 i12.2 i23.1 i11 i23
    |16                         |i8.1 i21.1 i8 i12 i21
    |                           |i6_____________________________proficient
0   |13                        X|
    |12                         |
    |11                         |i13.1 i16.1 i20.1 i13 i16 i20___partially
    |10                         |i6.2 i15.1 i15
    |9                          |i2.1 i3.1 i2 i3
    |8                          |i12.1
-1  |7                          |i22.1 i22
    |6                          |i1.1 i17.1 i1 i17
    |5                          |i9.1 i9
    |4                          |
    |                           |i14.1 i14
    |3                          |i10.1 i10
-2  |2                          |
    |                           |
    |1                          |
    |                           |
    |0                          |
    |                           |
==============================================================================
```

Figure 13. Evaluating the SEM for a respondent located at .05 logits

In Figure 13, this partially proficient student at the upper end of the SEM would meet the proficient threshold and at the lower end would remain as a partially proficient student.

IRT Assumptions and Item Fit

Two underlying assumptions of the PCM presented earlier are that the items are locally independent and that the latent variable being measured by the test is largely unidimensional. Under local independence, the assumption is being made that one's response to an item is reflective of a respondent's proficiency and should not be affected by other factors such as speeded testing situations, responses to other items, or when students are given varying different exposure to the material being provided on the test. Local independence refers to the conditional independence notion expressed earlier where the probability of a response vector for a given respondent is calculated by taking the product of the probabilities for each item on a test. Considerable violations to local independence could lead to inflated reliability estimates (Thissen & Wainer, 2001).

Similar to local independence, unidimensionality refers to the assumption that only one variable needs to be specified in order for the conditional probabilities of a correct item response to be independent. An example of a violation of this assumption would be a case in which some items on a math test require considerable facility with verbal comprehension (e.g., word problems). In such a case, two variables (i.e., "reading comprehension" and "math ability") would need to be specified in order for the assumption of local independence to hold, and this would make the test multidimensional rather than unidimensional. Although small violations of unidimensionality can be tolerated by an IRT model, considerable violations of unidimensionality can lead to incorrect inferences about student performance proficiency (Walker & Beretvas, 2003).

Following the assessment of the test's unidimensionalility and local independence, a third area assessed relates to how well the PCM fits the data. This investigation was limited to evaluating how well the expected responses under a PCM fit the observed responses to each item in the data set. Under IRT, each item has a functional form or an item response function (IRF) which shows how changes in proficiency or thetas relate to changes in the probability of responding correctly for a given MC item, or for each response category of a CR item. Under a PCM, the slope of this form is assumed to be constant across items. That is, all items are assumed to equally discriminate across respondents. Under more complex IRT models, the shape of the IRF form can vary in steepness depending on how well an item discriminates across respondents of different proficiency. MC items could be modeled using more complex extensions of the Rasch model, which allow the discrimination parameter to vary between items (the two-parameter logistic model) or allow for both discrimination and guessing parameters to vary between items (the three-parameter logistic model). More complex extensions of the PCM, such as the Generalized Partial Credit Model (GPCM; Muraki, 1992) used by the CSAP tests, relaxes the constant discrimination assumption under the PCM, and allows the discrimination parameter or the slopes to vary across both CR items. As noted earlier, the PCM was selected to evaluate the interpretive argument in this chapter over more complex extensions since the theta estimates from the PCM are more closely aligned to how the district used and still uses raw scores to classify students into different proficiency categories. Under more complex IRT models, the rank order of students may substantially differ from the rank order of students using the raw scores, since these complex models assign higher theta values to students who respond correctly or earn full points on items with higher levels of discrimination.

One way to evaluate how well a chosen IRT model, in this case the PCM, fits the data is to evaluate the item fit statistics. In ConstructMap, item fit can be evaluated by either reviewing the infit or the outfit mean square statistics in conjunction with the t-statistic provided. These fit statistics measure the average over the difference between the observed score and the expected score for a given item and is similar in interpretation to the concept of a residual produced under a regression analysis. That is, a large fit statistic is analogous to a large residual or large difference between an observed and expected score. The difference between the infit and the outfit mean square statistic, is that the outfit mean square statistic is unweighted and therefore influenced by respondents located farther away from an item. As an unweighted statistic, the outfit mean square is always larger than the infit or weighted mean statistic. ConstructMap uses a common standard of .7 and 1.3 to identify misfitting items. Items with an outfit mean square statistic located outside of these boundaries are considered to be misfitting items. However, Wu and Adams (in press) note that these boundaries were developed based on simulations using considerably smaller sample sizes ($n < 100$) than the sample sizes considered in the present analysis and that tests taken by larger samples of respondents should display item fit statistics closer to 1.

Two sets of exploratory analyses were conducted to evaluate the extent to which the test is largely unidimensional and the extent to which the items are locally independent. The item fit statistics generated by ConstructMap were then reviewed to evaluate how well the PCM fit the data.

Assessing Unidimensionality

A principal components analysis (PCA) using the R package Polycor was conducted to evaluate the extent to which the test is largely unidimensional; i.e., the degree to which grade 8 math item responses reflect a single latent dimension. This approach is also used by CTB-McGraw-Hill to assess CSAP test dimensionality. The variance found in the correlation matrix between items on a given instrument is represented by eigenvalues. An eigenvalue reflects the variance (reflected in a covariance matrix) that is shared among correlated items. Under a PCA, the scaled eigenvalues (and the proportion of variance they represent) are evaluated to assess the degree to whether test items reflect one or more unique dimensions. One method for assessing dimensionality through eigenvalues is represented by a scree plot. The scree plot represents a plot of all eigenvalues and resembles a slope with points or "scree" at the base of the slope. A scree plot identifies the number of factors or dimensions underlying the data based on the number of eigenvalues located before the junction where the slope connects with the base of the scree.

A second standard used to identify the number of primary components is the "Kaiser rule" (Kaiser, 1960). Under this rule, component represented by eigenvalues greater than 1.00 are identified as candidate components underlying items. The eigenvalues produced under the PCA suggest the presence of one underlying dominant component with an eigenvalue of 5.6 and one other minor component with an eigenvalue of 1.5.

The proportion of variance represented by individual dimensions is calculated by taking an eigenvalue and dividing that by the sum of all eigenvalues extracted. Using the Kaiser rule, the sum of eigenvalues extracted equals 7.1. The proportion of variance explained for the first component with an eigenvalue of 5.6 equals 79%. Eigenvalues and the item variance they represent suggest that there appears to be one highly dominant component that accounts for more variability than all other components flagged using the Kaiser rule.

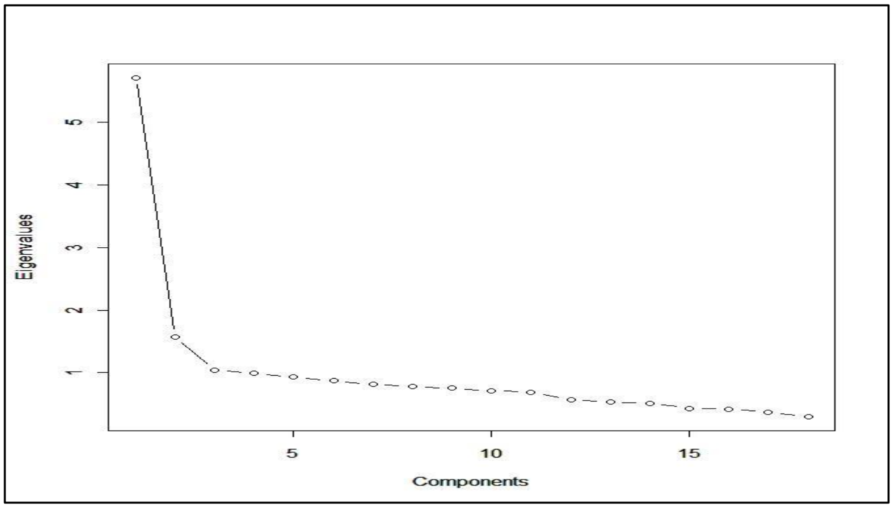


Figure 14. Scree plot for grade 8 math test

Figure 14 presents the scree plot generated for the grade 8 math test. Components located above the bend in the "elbow" in a scree plot signals the number of factors or components that appear to explain the most variability in responses across items. In Figure 14, the bend takes place between the second and third component and this finding suggests that although there may be two latent dimensions underlying this test, the first component is highly dominant and accounts for most of the variability found across items. That is, the test appears to be largely unidimensional.

Assessing Local Independence

The Q3 statistic (Yen, 1984) identifies the degree to which local dependence may be occurring between pairs of items on a test. The Q3 represents the correlation between the residual scores of the expected and observed values across all respondents for each pair of items. According to Yen (1984), the expected value under Q3 is equal to -1/(N-1), where N represents the total number of items on a given test When using larger samples of items, the Q3 would be expected to be closer to 0. Larger positive values (above .2) indicate that local dependence may be occurring between item pairs. Figure 15 shows the Q3 statistics between pairs of item on the grade 8 math test. Highlighted cells in the figure represent areas that show where local dependence may be occurring.

	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
I2	0.25																
I3	0.15	0.28															
I4	0.24	0.31	0.24														
I5	0.22	0.3	0.25	0.3													
I6	0.15	0.18	0.15	0.18	0.15												
I7	0.19	0.22	0.2	0.27	0.2	0.16											
I8	0.18	0.3	0.23	0.26	0.25	0.13	0.25										
I9	0.17	0.21	0.19	0.22	0.2	0.14	0.19	0.27									
I10	0.14	0.23	0.18	0.23	0.25	0.09	0.17	0.19	0.26								
I11	0.19	0.23	0.18	0.22	0.26	0.1	0.22	0.21	0.24	0.38							
I12	0.2	0.33	0.26	0.34	0.3	0.32	0.26	0.32	0.3	0.3	0.29						
I13	0.21	0.29	0.23	0.27	0.27	0.17	0.22	0.25	0.21	0.25	0.25	0.36					
I14	0.2	0.28	0.23	0.32	0.29	0.11	0.2	0.27	0.26	0.32	0.31	0.34	0.28				
I15	0.14	0.15	0.15	0.13	0.13	0.09	0.17	0.14	0.16	0.13	0.15	0.14	0.12	0.16			
I16	0.22	0.29	0.23	0.29	0.28	0.12	0.26	0.26	0.23	0.22	0.27	0.34	0.27	0.3	0.15		
I17	0.21	0.31	0.26	0.36	0.28	0.19	0.24	0.25	0.18	0.24	0.25	0.33	0.3	0.3	0.13	0.31	
I18	0.01	0.13	0.09	0.14	0.12	0.09	0.07	0.17	0.17	0.22	0.2	0.32	0.15	0.19	-0.03	0.21	0.14

Figure 15. Q3 statistic for paired items on the grade 8 math test

The large proportion of cells highlighted in Figure 15 suggests that there is a fair amount of local dependence occurring between pairs of items on the test. This finding would suggest that the SEM estimates would likely be biased downwards.

Assessing Item Fit

The outfit mean square statistic is an indicator of how well the model, in this case the PCM, fit the data. As noted earlier, a larger item fit statistic is indicative of a misfitting item. Wu and Adams (in press) note that a mean square fit statistic located considerably above 1 suggests that the item may not discriminate well between the responses of higher and lower proficiency students and a fit statistic below 1 typically means that the items are expected to have steeper curves and are expected to better discriminate between the performance of students with varying proficiency. Table 1 shows the outfit mean square statistic and accompanying t-statistic. Finding that the t-statistic for most items in the table is significant or located above 2 is expected since the t-statistic is highly influenced by sample size and a large number of students took this test.

Table 1

Item Fit Statistics for Grade 8 Math Test

	Item Outfit Mean Square Statistic	
Item	Outfit Unweighted Mean Square	T-Value
1	1.24	10
2	1.16	4
3	1.18	7.4
4	1.16	6.8
5	1.18	7.6
6	1.12	5
7	1.18	7.5
8	1.12	5.1
9	1.09	3.8
10	1.09	4
11	1.09	4.1
12	0.89	-5.2
13	1.14	0
14	1.1	4.5
15	1.22	9.3
16	1.08	3.6
17	1.16	0.5
18	1.01	3.8

Although the outfit mean square statistic for every item falls below the standard upper boundary of 1.3, the fit statistic based on having a large sample size (e.g., above 100 respondents) should be much closer to 1. Under common standards for test development, the fit statistics presented in Table 1 would be deemed as "acceptable", but based on Wu and Adam's (in press) research on item fit statistics, all values located farther away from 1 in Table 1 would suggest that the PCM does not fit the data very well.

On the basis of evaluating item fit, a common strategy for addressing the misfitting items would be to either discard the items being used or move to a more complex IRT model to fit the data (Embretson & Reise, 2000). Although the second option presents a viable strategy for addressing misfitting items on this test, using more complex IRT models would yield proficiency estimates that depart from the district's use of raw scores to drive decisions. Since the Rasch model is the only IRT model that yields proficiency estimates that correspond more closely to the raw scores (as depicted on the Wright Map), the Rasch approach serves as the best IRT option for modeling the data and to evaluate the extent to which the test may require more items to measure students of varying proficiency more reliably.

Based on the above set of analyses conducted to evaluate the two IRT assumptions and item fit, the evaluation of the interpretive argument which relies largely on findings from applying a PCM to one test was conducted with the understanding that since item fit is not perfect (though note that it would be considered acceptable under commonly used “rules of thumb”), the SEM may be slightly higher by the use of a PCM. However, the effects on the use of a PCM on the SEM may be considerably dampened due to the finding of local dependency occurring between items. Thissen and Wainer (2000) suggest the formation of testlets as one strategy for addressing local dependence. A testlet could be formed by bundling the scores across items that share local dependence. For example, if two items sharing local dependence are worth two points, the testlet would be formed by treating the two items as a single item with four possible points. In the case of the grade 8 math test, this testlet strategy would not be viable since there are few items on this test and according to Figure 15, most items on that test could be bundled together into a testlet. If testlets were formed on the grade 8 math test or with any other interim test, the smaller number of items left on the test would considerably lower the reliability of the instrument.

Since local dependency cannot be minimized for this test, the SEM will be biased downwards and this would subsequently have the opposite effect on the SEM from using a PCM to fit the data. The extent to which the effects of local dependency on the SEM cancels out the effects of misfitting items on the SEM are unknown, but understanding that both issues are present suggests that the SEMs reported in the findings should not be highly overstated (either too high or too low).

Interim Assessments Evaluated

The evaluation of the interpretive argument from the item design to the generalization inference in this chapter moves from an evaluation of a set of nine interim assessments from one grade in each level (elementary, middle and high) representing different content areas using classical item statistics to a more focused evaluation of the grade 8 math test using the PCM described previously. Table 2 below represents the set of assessments evaluated over the two-year time period in this study:

Table 2

Selected Interim Tests Assessed

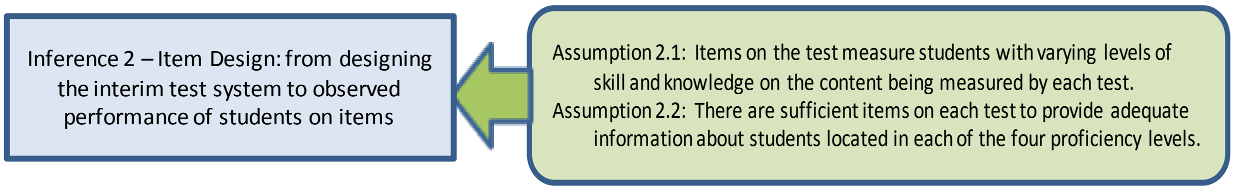
Grade	School Year	Test Version[a]	Content Areas	Testtakers (N)
4	2007-08	1	Reading	5023
4	2007-08	3	Reading	5075
4	2007-08	1	Math	5054
4	2007-08	3	Math	5084
8	2006-07	2	Math	3999
8	2006-07	2	Writing	4034
8	2007-08	2	Writing	4015
10	2006-07	2	Math	3224
10	2006-07	1	Reading	3297

[a] Test Version 1 = test administered in the beginning of the school year; 2 = test administered at the end of the first semester; 3 = test administered at the end of the school year.

The selected tests represent mathematics, reading, and writing; however, only the grade 8 math test was evaluated to empirically test the assumptions specified in the interpretive argument. This test was chosen since it met two of the three uses evaluated in this dissertation study. As noted in Chapter 2, all of the interim assessments were developed creating the same set of procedures and processes and were developed with the same intent of capturing the "power" standards measured and weighted more prominently on the CSAP.

All student level interim assessment and CSAP data from the 2006-07 and 2007-08 years used to evaluate the interpretive argument was provided by the DPS Assessment Department. The data from both interim assessments and CSAP consisted of the scored responses to each item. Individual student information was masked by Assessment Department to ensure that all individual student data remain secure.

Inference 2: Item Design



In this section, two assumptions are checked to evaluate the item design inference component of the interpretive argument: whether each test contains items assessing students with varying levels of skill and knowledge on the content being measured by each test, and whether there are sufficient items to evaluate students at varying levels of proficiency. P-values are used to evaluate whether the test has a variety of MC items ranging from easy to difficult to determine whether the first assumption supporting the item design inference holds. The p-values simply communicate the proportion of students who answered each MC item correctly. The items were selected by the panel to measure grade level skills and knowledge on state standards, and the extent to which the items can be answered correctly by students of varying proficiency on those standards is illuminated by the proportion of students responding correctly to each MC item. Since the p-values can only provide information on whether the test is populated with items that vary in difficulty, the second assumption is tested using the PCM for the grade 8 math test to evaluate whether there are sufficient number of items to reliably assess students at different proficiency locations. Table 3 shows the minimum, maximum, median and standard deviation of p-values for each of the nine tests reviewed. The p-value simply reflects the proportion correct for each MC item. Table 3 shows that with the exception of two tests, the minimum p-value typically ranged between .22 and .29. The maximum p-values presented in Table 3 also indicate that with the exception of the grade 10 math and reading tests, there was at least one item on each interim test where more than 70 percent of all students taking a given test could respond to correctly. The median p-values of three tests (grade 4 math, version 1; grade 4 reading, version 1, and grade 8 writing, version 2, 2006-07) indicate that these tests are easier than the others.

Table 3

P-Values for Nine Interim Tests

Grade	Year	Version	Subject	Minimum	Maximum	Median	Standard Deviation
4	2007-08	1	Math	0.15	0.88	0.67	0.19
4	2007-08	3	Math	0.22	0.87	0.51	0.17
4	2007-08	1	Reading	0.29	0.86	0.61	0.16
4	2007-08	3	Reading	0.25	0.82	0.50	0.15
8	2006-07	2	Math	0.27	0.81	0.41	0.15
8	2006-07	2	Writing	0.29	0.80	0.64	0.15
8	2007-08	2	Writing	0.13	0.72	0.41	0.19
10	2006-07	1	Math	0.26	0.65	0.35	0.12
10	2006-07	2	Reading	0.26	0.69	0.45	0.12

The grade 4 math and grade 8 writing tests fall outside the upper range of average p-values (.62) deemed as desirable under the standards applied to the CSAP tests measuring the same target domain. For the two easier grade 4 tests with median p-values of .67 and .61, these tests reflect the decision made by DPS staff to ensure that all version 1 tests are easier than their versions 2 and 3 counterparts. For the third easy test, the grade 8 writing test administered at the end of the first semester in 2006-07, this earlier version was updated with more difficult items when administered one year later. The bar charts in Appendix C-1 shows the p-value distribution for all nine interim tests.

In Appendix C-1, the distribution of p-values for MC items show that although each test consists of a few difficult items (.3 and below) and a few easy items (.7 and above), the majority of the items range in difficulty between .4 and .6. As expected, compared to the version 2 and 3 tests, the easier tests (version 1 tests) are populated with easier items and this would suggest that more students earned higher total scores on these tests relative to the second and third versions of the same test. In addition, the adjustment made by DPS staff to the version 2 test for grade 8 writing can be easily seen in Appendix C-1. Compared to the same version writing 2006-07 test, the 2007-08 grade 8 writing test was re-designed to reflect more difficult items as specified in the blueprints. It is worth noting here that under the use of these tests for professional compensation purposes (evaluated in Chapter 5), that the easier version 1 tests were used as a baseline for determining how many students were proficient compared to the more difficult version 3 test. The pattern of finding easier items for the version 1 test relative to the version 3 test holds for grade 4 reading and math (see Figure 9), and according to DPS assessment staff, should hold for all other tests. For merit pay, many teachers were encouraged to develop their growth objectives based on moving more students to proficiency on the version 3 tests. In addition, teachers used results from the version 2 tests in the 2006-07 and the 2007-08 years to revise their growth objectives. Since the raw scores were and are currently being used for devising growth objectives using the interim tests, differences in test difficulty have implications for fairly evaluating growth gains being made by students between versions. The assumptions underlying the use of these tests specifically for merit pay purposes are evaluated in the next chapter.

Although most tests (with the exception of version 1 test and the older version 2 grade 8 writing test) have items covering a range of difficulty, the p-value distributions presented in Figure 10 do not provide information about whether there are sufficient items developed for each proficiency level to make a good determination of a student's proficiency. That is, whether there are an adequate number of items measuring students at varying proficiency levels to give decision makers good information to drive decisions. As discussed previously, to evaluate whether sufficient items at different points along the difficulty range are available, the IRT generated Wright Map is evaluated along with the conditional SEM.

Similar to the finding of an average p-value for this test of .45, the Wright Map for this test shown in Figure 16 suggests that on average, more than half of all respondents were estimated at a location below approximately half of all items on this test. All respondents below 0 logits had less than a 50% chance for endorsing approximately half of all items used on this test.

```
       Map of person estimates and response model parameter estimates (Note:  each “x” represents
34 students, each row is .255 logits)
=========================================================================
   raw      students                          item.step
-------------------------------------------------------------------------
3  |                       |
   |                       |
   |                       |
   |24                    X|
2  |23                    X|CRi6.1 CRi12.3
   |22                    X|CRi6.2 CRi18.4 _____________________________________Advanced
   |                     XX|
   |19                   XX|
1  |                 XXXXXX|
   |16                 XXXX|i3.1 i7.1_______________________________Proficient
   |            XXXXXXXXXXX|CRi12.2 i15.1 i17.1
0  |           XXXXXXXXXXXX|i1.1 i2.1 i4.1 i8.1 i13.1 CRi18.3 ____________Partially Proficient
   |      XXXXXXXXXXXXXXXXX|i5.1 i16.1 CRi18.1 CRi18.2
   |9  ----------XXXXXXXXXX|
   |   XXXXXXXXXXXXXXXXXXXX|i9.1 CRi12.1 i14.1
-1 |6          XXXXXXXXXX|
   |      XXXXXXXXXXXXXXXXX|
   |3                XXXXXX|i11.1
   |2                   XXX|
-2 |1                    XX|i10.1
   |0                     X|
   |                       |
```

Figure 16. Wright map of person estimates and response model parameter estimates grade 8, version 2 math 2006-07

As seen in Figure 16, the majority of items in the Wright Map appear to be clustered just above the mean of the respondent distribution noted by the dotted line, with a few items that are located at either end of the respondent distribution. All items contribute to the estimation of a respondent’s location, but items located closer to a respondent contribute better information about that respondent’s location than items located farther away from a respondent. The Wright Map presented in Figure 16 shows that the items on the grade 8 math test should ideally provide more reliable estimates of students located in the partially proficient region since more items are contributing to the estimation of those students. This initial observation of the Wright Map in Figure 16 would indicate that one would expect the standard error of measurement (SEM) associated with scores located at around the proficient cut to be smaller than the SEM for scores located farther away from the cluster of items.

Although reliability associated with the SEM estimates are evaluated more closely under the generalization inference, the SEM is discussed briefly here to evaluate whether there are sufficient items to make judgments about respondents at two critical points of the Wright Map: the cut-point between “unsatisfactory” and “partially proficient”, and the bound between “partially proficient” and “proficient”.. As noted in earlier chapters, the unsatisfactory cut for this test determined which individuals required mandatory summer remediation during the summer of 2007. The proficiency cut for this test provided teachers with information on whether they should revise their student growth objectives based on how many students became proficient between the first interim and second interim assessment, or identified students who may need additional tutoring or remediation prior to the CSAP testing in February (for grade 3 students) or March (for all other grades). Figure 17 presents the SEM for the grade 8 math test. The SEM curve, in contrast to the curve presented earlier for another interim test, appears to be relatively flat across respondents.

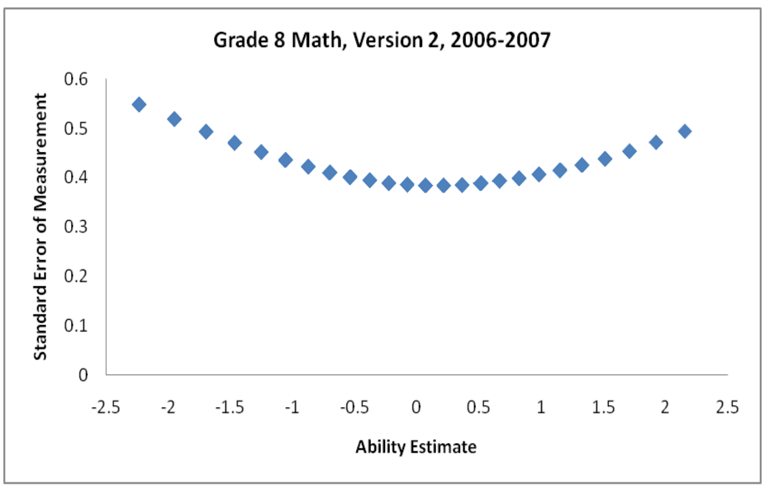


Figure 17. SEM for grade 8 math test

This finding indicates although items tend to cluster near the the proficient cut, students located at varying proficiency levels are assessed with a similar degree of error of approximately .4 logits. Although finding that the location of most students are estimated with a similar degree of error is not in itself a problem and may even be ideal for the purpose of meeting the needs of teachers interested in acquiring information about students from all proficiency levels, the decision to establish four different proficiency areas based on a small set of items appears problematic. As pointed out in Figure 17, the partially proficient band is narrow and the unsatisfactory score range spans almost half of the entire range of abilities measured by this test. When considering the SEM relative to the small range used to classify partially proficient students, the ability to distinguish a partially proficient from an unsatisfactory and a proficient student becomes more uncertain.

To illustrate the uncertainty of identifying a partially proficient student, the SEM was examined relative to the estimated locations for two different students-- “Ed” and “Jon” -- with test scores that would place them respectively into the partially proficient and unsatisfactory categories. Ed, the partially proficient student receives a score estimated to be at .21 logits, whereas Jon’s estimated score is located at -.4 logits. The two blue circles in Figure 18 shows the estimated location of Ed’s and Jon’s scores on the Wright Map, with the partially proficient score of Ed located higher on the map.

```
=========================================================================
   raw      students                              item.step
-------------------------------------------------------------------------
3   |                        |
    |                        |
    |                        |
    |24                      |
2   |23                      |i6.1 i12.3
    |22                      |i6.2 i18.4
    |                        |
    |19                      |
1   |                        |
    |16                      |i3.1 i7.1  _______________________________Proficient
    |                        |i12.2 i15.1 i17.1
0   |                        |i1.1 i2.1 i4.1 i8.1 i13.1 i18.3 ________________Partially Proficient
    |                        |i5.1 i16.1 i18.1 i18.2
    |9                       |
    |                        |i9.1 i12.1 i14.1
-1  |6                       |
    |                        |
    |3                       |i11.1
    |2                       |
-2  |1                       |i10.1
    |0                     --|
```

Figure 18. Wright Map showing estimated locations of Ed and Jon

The red bars located at either end of each blue circle represent approximate lengths of the standard error associated with the location of each respondent’s estimated score. For Ed, the upper bound estimate of the SEM places his score in the proficient region. The lower bound estimate almost places Ed’s score in the unsatisfactory range, indicating that all partially proficient students with scores slightly lower than Ed's could also be classified as unsatisfactory students. The SEM associated with Ed's score is .38 logits and the SEM estimate associated with Jon's is .39 logits. The information gleaned from evaluating the SEM associated with Ed's score indicates that partially proficient students below him would have a SEM that would classify them as unsatisfactory and partially proficient students above him could be estimated as falling in the proficient region. In the case of Jon, the upper bound estimate places his score in the partially proficient area.

The question of whether the SEM of .4 is too large or is acceptable depends largely on the uses of the test. According to the test developer of the CSAP, a desirable quality of a test using cut-points is if the SEM located at the cut point for proficiency is smaller than the SEMs associated with other points of the scale. Because a substantially smaller set of interim assessment items are administered to students, the SEM on these interim tests would naturally be larger than the SEMs found on a large test such as the CSAP. Relative to the CSAP, the grade 8 SEM associated with the proficient cut is equal to .4 compared to .2 on the CSAP tests. This finding would suggest, that compared to the CSAP, a considerably larger range of partially proficient students could by chance belong to the proficiency category. A strategy for lowering the SEM at the proficiency cut would be to add more items on the test to provide even better estimates of students located at that region or to shift towards a computer adaptive testing (CAT) system. However, since neither option is viable at this time due to high cost implications, an alternative strategy to consider would be to only use the test to determine students located at an above proficient or below proficient cut since there are not enough items to properly distinguish four different proficiency regions. Although this would have implications for driving decisions or test uses, this strategy would address the uncertainty associated with reliably classifying all partially proficient students.

The question of whether the SEM of .4 is too large or is acceptable depends largely on the uses of the test. According to the test developer of the CSAP, a desirable quality of a test using cut-points is if the SEM located at the cut point for proficiency is smaller than the SEMs associated with other points of the scale. Because a substantially smaller set of interim assessment items are administered to students,

the SEM on these interim tests would naturally be larger than the SEMs found on a large test such as the CSAP. Relative to the CSAP, the grade 8 SEM associated with the proficient cut is twice as large at .4 compared to .2 on the CSAP tests (CDE, 2009[16].). A strategy for lowering the SEM at the proficiency cut would be to add more items on the test to provide even better estimates of students located at that region or to shift towards a computer adaptive testing (CAT) system. However, since neither option is viable at this time due to high cost implications, an alternative strategy to consider would be to only use the test to determine students located at an above proficient or below proficient cut since there are not enough items to properly distinguish four different proficiency regions. Although this would have implications for driving decisions or test uses, this strategy would address the uncertainty associated with reliably classifying the performance of many students.

Returning to the examples of Ed and Jon, despite their differences in proficiency classifications, the SEMs for both students are large enough to make it difficult to know whether they may by chance belong to one proficiency category or to another. Within the context of using this test for predictive purposes and for identifying unsatisfactory students for remediation, the large SEM associated with Ed and Jon's score had different implications for each student. In Jon's case, although the SEM associated with his score indicates that he could have been classified as a "partially proficient" student, his classification as an "unsatisfactory student" was associated with negative consequences for him. As an "unsatisfactory" student, his parents were notified by the district that he was not ready to transition into high school math classes, and he was required to attend a summer remediation academy. In Ed's case, a teacher may have decided to offer additional tutoring or interventions to better prepare him for the CSAP tests, and the consequences associated with this predictive use may have benefitted him (unless the additional services require that he forfeit participation in an elective to attend CSAP workshops).

Although both students may benefit from remediation (for Jon) or extra tutoring (for Ed), the large SEMs call into question whether these actions were warranted based on information gleaned solely

[16] The CSAP technical manual provides the proficient cut point and associated SEM for the grade 8 math test on the scale score metric. This scale score values were transformed into logits using the linking constants provided in the technical manual.

from this one test. The same issue regarding the level of uncertainty relative to having four proficiency level defined on a test with fewer than 30 items applies to the other interim tests. The plots in Figure 19 present the SEM found for four other interim tests.

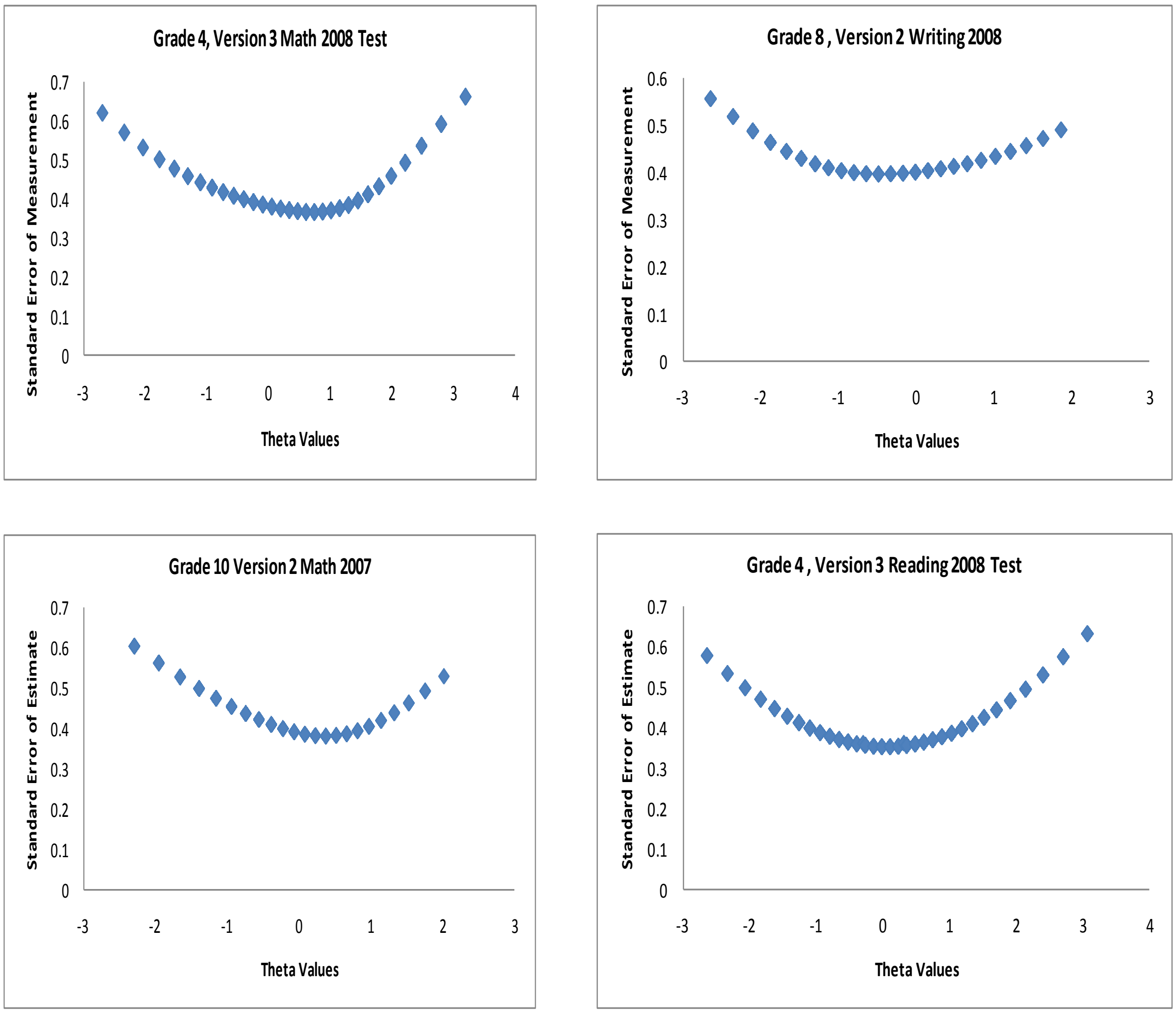
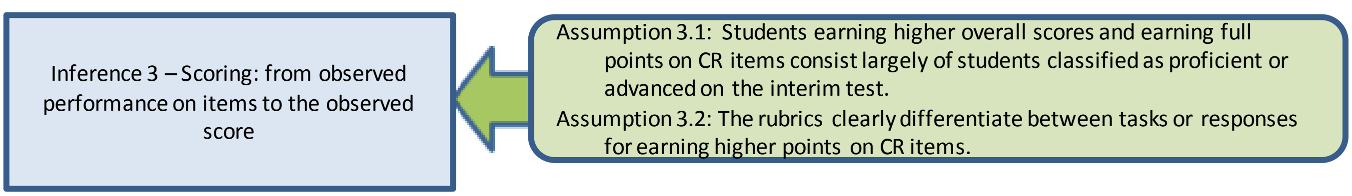


Figure 19. SEM plots for other interim assessments

For all four tests, the minimum SEM values sit above .35. Because a small range of scores apply to each proficiency category, a number of students in each region could be classified by chance as in either a higher or a lower proficiency band when factoring in the distance of the SEM.

The findings presented under item design foreshadow the connection between item design and the generalization inference. That is, making reliable judgments about proficiency based on these interim tests with 25 or fewer items can be problematic for driving uses requiring absolute decisions. The extent to which students fall under one or more proficiency categories for the grade 8 math test is discussed in more detail under the generalization inference section. Before evaluating the assumptions under the generalization inference, however, the next section evaluates whether the scores earned by students on individual items appear to be consistent with their proficiency classifications. The assumption checks for the scoring inference serve to establish whether ratings, particularly on the CR items that require students to demonstrate or show their work, appear to be fairly consistent with the performance level attained. For example, if an "advanced" student scores low on a given CR item, it would be expected that the item is very difficult and students classified in lower categories would not be rated as scoring higher than advanced students. A finding that reveals otherwise may highlight possible problems with how the scoring rules are interpreted or implemented by raters.

Inference 3: Scoring



As noted earlier in Chapter 3, the scoring inference for the validation study pursued in this dissertation focuses on two areas that can be evaluated using the available data: whether higher points earned on constructed-response (CR) items scored by raters are earned largely by students with higher levels of proficiency, and whether the rubrics provide clear guidance to raters evaluating each student's work on constructed responses. Although a rater scores each interim test, the raters are only asked to exercise judgment over scoring CR items. Since three to four CR items used on each interim test accounts for a third of all possible points that a student could earn on a test, this chapter focuses primarily on evaluating the quality of the CR items on the grade 8 math test using item-total correlations. As detailed earlier in this chapter, the grade 8 math test, which meets two of the three uses evaluated in this chapter, was used largely to test out the assumptions supporting the scoring inference and all other inferences evaluated in this chapter.

Although the first assumption supporting the scoring inference is not reflective of a commonly held assumption by test developers, it is important to recall from Chapter 2 that the district instructed item panel members to assess each response category relative to their conceptions about student proficiency. In a document used to guide panel members in their efforts to modify and refine items for each interim test, item members were tasked with the exercise of "writing in the number of points (for each dimension on the rubric) you think a student at each level of proficiency would achieve". The task of identifying CR response categories relative to conceptions about proficiency applied to every item on all interim tests. The idea behind this task was to ensure that each test had an adequate number of items evaluating students considered to be below or above proficient. Again, although the first assumption supporting the scoring inference reflects an assumption about the district's expectation on the behavior of each CR item, this assumption may not necessarily hold if considering other factors that would enable below proficient students to earn full points on a CR item. That is, some below proficient students could perform better on CR items that afford them the opportunity to write out their responses, and may perform better on the CR item but score poorly across MC items. For the purpose of evaluating the scoring inference, the first assumption is evaluated to determine the extent to which these items conform to this district's expectations and rules established to ensure that the performance of proficient and above students is distinguished from the performance of below proficient students on CR items.

The scoring rubrics accompanying each CR item on the grade 8 math test were also reviewed to check on whether the second assumption supporting the scoring inference holds. That is, if the guidelines provided in the rubrics appear to provide clear information establishing what students need to accomplish in order to earn higher points, this finding may suggest that unexpected scores earned by students may point to other factors. These factors may include raters who may not be following the rubric guidelines, content that may not have been taught at the time of the test administration, or students who may not be interpreting or understanding the prompt or the language used in the item.

Prior to reviewing the CR items for the grade 8 math test, the item-total correlations expressed by point-biserials were first reviewed for all MC and CR items on each of the nine interim tests. The point-biserial represents the correlation between scores on each dichotomous or response category (for each CR item) and the total scores. For an MC item, a low point-biserial indicates that the item does not appear to discriminate well between students with higher total scores who earn full points on the item relative to students with lower total scores. The scatter plots presented in

Figure 20 provides a basis for understanding the difference between an item with a point-biserial lower than the standard of .3 relative to an item with a point-biserial well above the standard at .5.

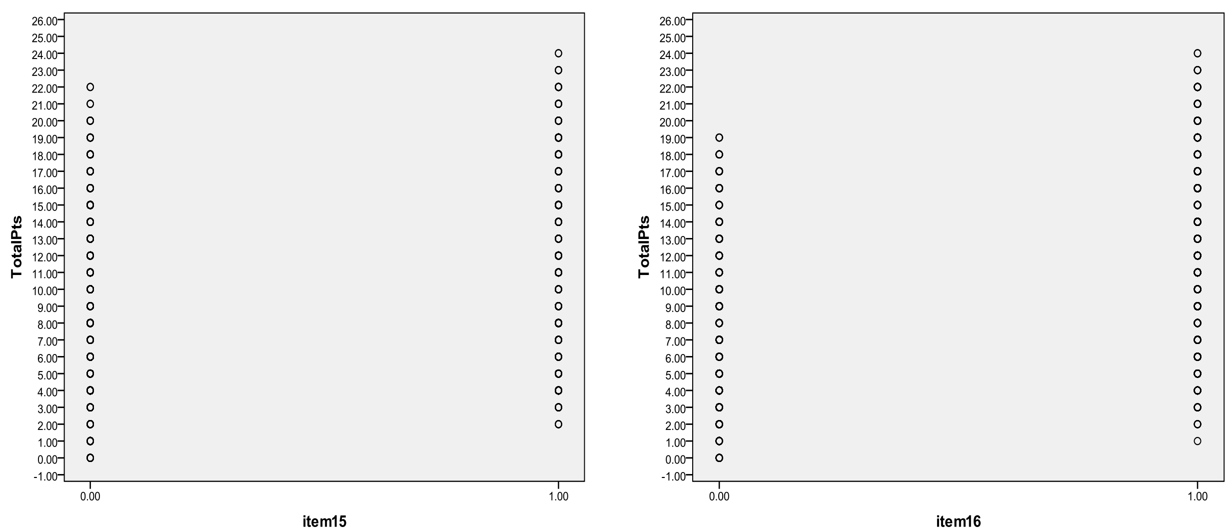


Figure 20. Comparison of items with high and low point-biserials

In Figure 20, Item 15 has a point-biserial of approximately .2 and item 16 has a point-biserial of approximately .5. The scatter plot for item 15 shows that there appears to be less of a distinction between higher and lower performing students who score either a 0 or a 1 on this item. In contrast, the scatter plot for item 16 shows a more distinct difference between how students with higher and lower total scores perform on this item. The scatter plots for these two items show that as point-biserial correlation increases from .2 to .5, the item improves in its ability to distinguish the performance between students with lower and higher total scores.

For CR items, the higher response categories for each CR item should ideally correspond with higher point-biserials. Further, the highest response category should also reflect a point-biserial located above the .3 standard since the highest response category of a CR item was expected (at least by DPS staff expectations) to differentiate between the performance of proficient and above and below proficient students.

Before evaluating the CR items, a preliminary check of MC items across all nine tests was conducted to evaluate the extent to which scores on each MC item appear to correspond to expectations of how lower and higher proficiency students would perform. Using the same standard as the CSAP, MC items with point-biserials below .3 were flagged as items that may not be providing useful information to distinguish the performance of higher and lower proficiency students. Table 4 presents a summary of the point-biserials found on all MC items across all nine tests.

Table 4

Median Point-Biserials for Selected Interim Tests

Grade	Year	Version	Subject	Median Point-Biserial	Standard Deviation	Minimum	Maximum
4	2007-08	1	Math	0.46	0.07	0.21	0.55
4	2007-08	3	Math	0.44	0.09	0.3	0.57
4	2007-08	1	Reading	0.47	0.08	0.23	0.59
4	2007-08	3	Reading	0.44	0.09	0.24	0.57
8	2006-07	2	Math	0.45	0.08	0.18	0.49
8	2006-07	2	Writing	0.51	0.06	0.34	0.6
8	2007-08	2	Writing	0.39	0.15	0.05	0.58
10	2006-07	1	Math	0.45	0.09	0.21	0.55
10	2006-07	2	Reading	0.40	0.11	0.1	0.57

The median[17] point-biserials provide an indication of whether all items, on average, tend to discriminate between the performance of students who score higher or lower on the test overall. The standard used by test companies and applied to these tests is a mean (in this case median) point-biserial of .3. As indicated earlier, if a median point-biserial for a test is found to be lower than .3, this serves as a signal to test developers that the assessment may include items that are providing useful information about students.

[17] The median is reported in the table since this measure is less sensitive to extreme values

In Table 4, the minimum point-biserials suggest that with the exception of the grade 8 version 2, 2006-07 writing test, there was at least one item that was located below the .3 standard typically used by test developers as a criterion for re-evaluating or eliminating an item. The minimum point-biserial presented for each of the nine tests show that all but one test had at least one item with a point-biserial located below .3. The distribution of point-biserials for each test is presented in Appendix C-2. Overall, the bar charts in Appendix C-2 reveals that with the exception of one to three MC items on each test, most items have point-biserials indicating that they are able to differentiate between the performance of lower and higher proficiency students. Typically, during a test development process, items exhibiting correlations lower than .3 would be discarded since these items do not contribute good information about students or does not improve the reliability of the test.

In the case of the DPS assessment system, since the interim tests contain relatively few items, improving items – particularly CR items carrying more scoring weight than MC items would result in improving the reliability of the test. For example, Table 5 below presents the point-biserials for each CR item and Cronbach's Alpha if the CR item was not included in the grade 8 math test. When considering all 18 items of the test together, the assessment has an alpha level of .75.

Table 5

Alpha and Item-Total Correlations for CR Items on Grade 8 Math Test

Item	Item-Total Correlation[a]	Cronbach's Alpha if Item Deleted
6	.26	.74
12	.43	.71
18	.39	.73

Table 5 represents the point-biserial between the highest point-biserial found for each item and the total scores for all three CR items on the grade 8 math test. For item 6, the point-biserial of .26 is just below the standard of .3, and the last column of the table reveals that if this CR item were removed from the test, this would have little effect of lowering the original alpha level of .75. In other words, this CR item does not appear to contribute much in the assessment of student performance on this test. Items 12 and 18 represent items with slightly higher point-biserials that can better discriminate between the performance of students of higher and lower proficiency. In the case of item 12, the last column of Table 5 indicates that reliability drops substantially from .75 to .71 if this item is removed from the pool. For item 18, the reliability drops from .75 to .73 if this item is not included in the overall pool of items.

The tables in Appendix C-3 present the point-biserials associated with each CR response category for each of the nine tests. The point-biserial tables displayed in Appendix C-3 highlight an issue occurring with at least one CR item on five out of nine tests. For each of those five tests, there was at least one CR item exhibiting a low point-biserial at the highest response category or a decreasing point-biserial at the highest response category. CR items displaying undesirable qualities as flagged by the point-biserials were located in grades 4, 8 and 10 in reading, writing, and math. Considering that eight out of nine interim tests have one or more MC items with point-biserials below .3 and that five out of nine tests in both years of this study and across different grades and content have CR items flagged, the information in Table 5 suggests that those items should be re-evaluated to improve the overall quality of the test. The set of analyses that follows provides an example of the different steps that may be taken to evaluate CR items displaying undesirable qualities and identify areas where these items could be improved.

The following steps were taken to identify some of the probable causes for why point-biserials for some CR items were not increasing in value for each successive step or why the point-biserials did not reach an ideal level (.3 and above) for the highest response category:

1. Crosstabs comparing total scores and proficiency classifications earned relative to points earned on CR items were reviewed. This information provided an initial scan of how many students were located in each scoring category and whether those students appeared to correspond to the item panel's expectation of seeing the majority of proficient/advanced students earn full credit for each CR item.

2. A math content specialist was asked to review each item to determine whether the item prompt was clear and whether the content was reflective of grade level expectations.

3. The scoring rubrics were reviewed to determine whether the scoring rules provided clear guidelines for rating student responses.

Table 6 displays the point-biserials for the grade 8 math test which consists of 15 MC items and 3 CR items. As indicated by Table 6, all three of the CR items highlighted displayed unexpected qualities.

Table 6

Point –Biserials for grade 8 math test

	Response Categories				
Item	0	1	2	3	4
1	-0.3	0.3			
2	-0.47	0.47			
3	-0.37	0.37			
4	-0.48	0.48			
5	-0.45	0.45			
6	-0.24	0.12	0.26		
7	-0.37	0.37			
8	-0.44	0.44			
9	-0.4	0.4			
10	-0.47	0.47			
11	-0.45	0.45			
12	-0.52	0	0.43	0.31	
13	-0.46	0.46			
14	-0.49	0.49			
15	-0.18	0.18			
16	-0.48	0.48			
17	-0.47	0.47			
18	-0.55	-0.08	0.19	0.39	0.34

The first CR item on the test, item 6, has a point-biserial below .3 at the highest response level. This finding suggests that the highest response category for item 6 did not appear to differentiate between performance of higher from lower scoring students as expected. To evaluate whether the majority of proficient and advanced students earned full points on this item as anticipated, the relationship between a student's proficiency level earned on this test relative to the score earned on this item was assessed.

Table 7

Points earned by proficiency classification on grade 8 math test for Item 6

	Proficiency Level Score Range and Percentage of Students							
	A[a]	A	P	P	PP	PP	U	U
Response Category	22-24	%	17-21	%	12-16	%	0-11	%
Earned 0 points	1	5%	286	69%	608	82%	2394	91%
Earned 1 point	4	21%	67	16%	101	14%	220	8%
Earned 2 points	14	74%	63	15%	32	4%	24	1%
Total	19	100%	416	100%	741	100%	2638	100%

[a]A = advanced; P = Proficient; PP = Partially Proficient; U = Unsatisfactory.

In Table 7, the score range under each proficiency classification is based on the same cut-points applied by DPS staff across all interim tests in 2006-07 for every subject and grade area. In the previous section of this chapter (item design inference), item 6 was highlighted on the Wright Map as an extremely difficult item where the likelihood of higher proficient students earning either 1 or 2 points on this item is less than 50% and the likelihood of partially proficient and unsatisfactory students earning 1 or 2 points on this items is substantially lower than 50%. Overall, although the item panel's expectations of having more "advanced" students earn full credit and having the majority of below proficient students scoring 0 points on this item were fulfilled, Table 7 indicates that the proficient students' response pattern deviated from the expected beahivor for this CR item. As indicated by the table, the majority (69%) of students classified as "proficient" on this test earned 0 points on this item, with only 15% of all proficient students earning full points.

According to the DPS curriculum specialist, the finding of not seeing many proficient students earning partial-credit or full points on this item would not be surprising to grade 8 math teachers since item 6 reflects a math problem that only a small proportion of grade 8 students could attempt. Item 6, presented below, shows a graphing task requiring familiarity with algebra:

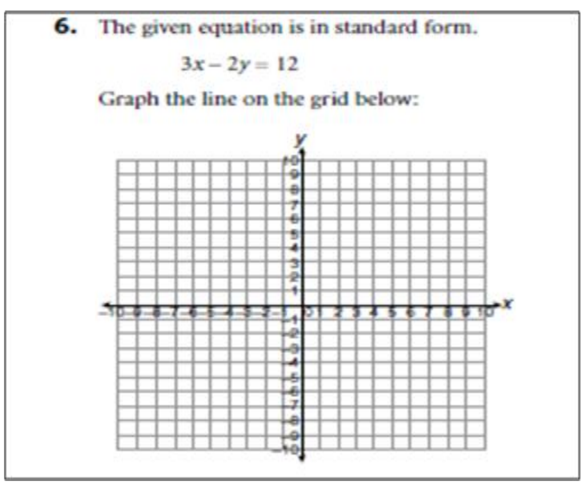


Figure 21. Item 6
(Reprinted with permission from the Denver Public Schools)

Although algebraic concepts are explored in the standard grade 8 curriculum, algebra is not formally studied until grade 9. The curriculum specialist also indicated that for this particular test, a few items testing grade 9 Algebra I content were deliberately included to screen students who wanted to skip Algebra I in high school. According to district staff, the use of the grade 8 math test to fulfill this secondary purpose was also continued into the 2007-08 school year.

The next step of evaluating this item entailed a review of the holistic rubric used by raters to score this item. Based on guidelines provided in the scoring rubric, it seems unlikely that the scoring rules contributed to the outcome of finding fewer proficient students scoring higher points on this item. The following rules provided to teachers clearly differentiate the level of skills or tasks required to move from one response category to another:

2 points: Responses include a coordinate graph with data plotted correctly.

1 point: Responses include a coordinate graph with most data plotted correctly, but contains a plotting error (single data point, *y*-intercept or slope).

0 points: Responses demonstrate no evidence that student has mathematical knowledge of constructing a line graph from a given equation of a line.

Considering that the scoring rules in the rubric appear to clearly distinguish the work required to earn higher points on this item, it seems unlikely that these rules misled raters to give a large proportion of proficient students zero points. For item 6, it would seem that the theory posited by the curriculum specialist appears more plausible for supporting the finding that fewer proficient students did not earn partial or full-credit as anticipated.

The next two items evaluated, items 12 and 18 represent CR items with point-biserial correlations that decrease in the highest response category. Wilson (2005) notes that point-biserials should ideally increase from low to high with each successive step or higher response category; and that higher point-biserials associated with higher steps indicate that higher and lower ability students are becoming more differentiated in their performance. Although the uncharacteristic decrease in item-total correlations cannot be attributed to any one factor, the same steps used to evaluate item 6 were also used for these items to determine whether any problems could be detected with either patterns in raw responses relative to proficiency classifications, the content or wording of the item, or with the scoring rubric used by raters to score the responses.

Table 8 presents the distribution of respondents in each scoring category for item 12 relative to total scores earned on the grade 8 math test and corresponding proficiency classification. In general, the pattern of responses based on proficiency classification in Table 8 shows that although students in the advanced and unsatisfactory categories fulfilled the design expectations for this item, the item does not seem to clearly differentiate between the performance of partially proficient and proficient students. As indicated by Table 8, the majority of "proficient" students earned 2 rather than 3 points for this item, and that the majority of "partially proficient" students also earned 2 points for this item.

Table 8

Points earned by proficiency classification on grade 8 math test for Item 12

	Proficiency Level – Score Range and Percentage of Students							
	A[a]	A	P	P	PP	PP	U	U
Response Category	22-24	%	17-21	%	12-16	%	0-11	%
Earned 0 points	0	0%	12	3%	68	9%	1282	49%
Earned 1 point	0	0%	85	20%	258	35%	993	38%
Earned 2 points	8	42%	227	55%	341	46%	322	12%
Earned 3 points	11	58%	92	22%	74	10%	41	2%
Total	19	100%	416	100%	741	100%	2638	100%

[a]A = advanced; P = Proficient; PP = Partially Proficient; U = Unsatisfactory.

The DPS curriculum specialist reviewing item 12 contended that the item may not be displaying a favorable quality since the wording of the item prompt could be misleading to some students. Item 12 is presented in the following page.

12. A group of scientists studied salmon in a mountain stream. They collected the following data by catching and releasing fish.

Length (in inches)
15
14
17
14
22
24
16
15

Part A Find the median and the mode of the set of data. Clearly label each answer and show any work in the space below.

Part B The mean is equal to 17.125. Which measure of central tendency describes the data the best? Explain your thinking.

Figure 22. Item 12
(Reprinted with permission from the Denver Public Schools)

In the first part of the problem, full credit for the item is only given if students explain how they found the median and the mode of the data set. The math specialist indicated that many students in grade 8 should easily identify the median and mode but a few students located at any level of proficiency may have overlooked the other task of explaining the logic for deriving both measures of central tendency. Another issue pointed out by the specialist about the item is that the question implies that there is only one mode to the data set when the data table presents two modes. The wording of the prompt could potentially confuse a student asked to identify a single mode.

For the second part of item 12, students are asked to identify, "…which measure of central tendency describes the data best? Explain your thinking." According to the DPS math specialist, the wording of the second part of the CR item could also confuse students who would have likely responded correctly if the question asked whether the mean, median or the mode best describes the data set. In addition, the specialist indicated that not all students can grasp the nuances or distinction between each measure if some students have not yet learned or fully mastered these concepts in class by the time they took this mid-year assessment.

The scoring rubric was also reviewed to see whether the rules for earning two points or one point on this item are clearly specified. The rubric for item 12 states:

3 points: Responses include 2 correct responses and work that shows how the student arrived at the answers. (Part A: Median = 15.5. Modes = 14 and 15). Part B: Median would be the best measure of central tendency to describe the data set which has skewed distribution (outliers on one side of the distribution).

2 points: Responses include 2 correct answers but no work showing how student arrived at those answers. Responses include 1 correct answer with work shown, and 1 incorrect answer with work, or an explanation that shows that the student had some idea of how to calculate measures of central tendency, or to determine which measure of central tendency describes the data best.

1 point: Responses include 1 correct answer with work shown or an explanation and 1 incorrect answer without work shown or an explanation. A response that includes 2 incorrect answers but enough work shown to demonstrate that the student had some understanding of how to calculate measures of central tendency, or to determine which measure of central tendency describes the data best.

0 points: Responses demonstrate no evidence that student has mathematical knowledge of displaying and using measures of central tendency in problem-solving situations.

The rule for earning three points is clearly defined. According to the rubric, a response earning full points would show all parts of item 12 completed and show all work associated with deriving the median and the modes. However, the rubric rules presented to raters for assigning a rating of 2 or 1 are confusing and do not clearly distinguish the separate conditions required for each scoring category. For example, in a hypothetical case, an "unsatisfactory" student might respond to CR item 12 as follows:

Response to Part A: Writes down the median and explains how she calculated the median. Writes down one mode, but fails to explain how she found the mode.

Response to Part B: Writes down an incorrect answer about which measure she thinks best describes the data and provides an explanation of why she thinks the measure she chose best describes the data.

Based on the guidelines provided in the rubric, a rater scoring this hypothetical student's response could rate this response under a 1 or a 2. In order to earn two points, a student could write down the median and the modes and provide no supporting information on how she found each data point. Another rule for earning two points is that the student can write down one correct answer with supporting information and one incorrect answer with no supportive information, or provides “some idea of how to calculate measures of central tendency”. If the rater believes the student’s incorrect response to Part B shows some idea of how to calculate measures of central tendency, then based on fulfilling different conditions under the rules specified for earning two points, this rater could be led to believe that this response should be rated as a 2. Another rater, however, may decide that since this student’s response meets the first rule stated for earning one point, and may subsequently assign one point to this student.

Under a different hypothetical scenario, a higher performing “proficient” student may respond to the same item as follows and still earn two points:

Response to Part A: Writes down the median and the modes, but fails to explain how he found both answers.

Response to Part B: Provides a full explanation of which measure of central tendency best describes the data.

The above hypothetical responses ideally should be considered as demonstrating more mastery over the content being tested, in particular because the response to Part B contains a detailed explanation of why the median serves as the best measure of central tendency. However, based on the rubric rules, because the answer to Part A fulfills the first condition for earning two points, a rater could simply ignore the valuable information provided in Part B and determine that this response warrants two points. In the case of this item, the confusing set of rubric rules governing the first and second scoring categories could be largely responsible for seeing many partially proficient and proficient students earn two points on this item.

For the last CR item reviewed in this section, the pattern of responses relative to the proficiency classifications shows a similar outcome to the pattern found on CR item 12. Table 9 presents the number of students landing in each response category for item 18 by the total score range and proficiency classification earned on the grade 8 math test.

Table 9

Points earned by proficiency classification on grade 8 math test for Item 18

	Proficiency Level – Score Range and Percentage of Students							
	A[a]	A	P	P	PP	PP	U	U
Response Category	22-24	%	17-21	%	12-16	%	0-11	%
Earned 0 points	0	0%	5	1%	63	9%	1265	92%
Earned 1 point	0	0%	28	7%	145	21%	696	51%
Earned 2 points	0	0%	93	23%	214	32%	426	31%
Earned 3 points	6	32%	180	44%	244	36%	221	16%
Earned 4 points	13	68%	110	27%	75	11%	30	2%
Total	19	100%	411	100%	678	100%	1373	100%

[a]A = advanced; P = Proficient; PP = Partially Proficient; U = Unsatisfactory.

Similar to the pattern found for CR item 12 the response pattern generally conforms to expectations with most advanced students scoring higher points on this item and with the majority of unsatisfactory students scoring 0 points on this item. However, the table indicates that this item may not be differentiating the performance of partially proficient and proficient students since a large proportion of partially proficient and proficient students populate the 3 point response category.

The next step of evaluating this item drew on the expertise of the curriculum specialist. Item 18 presented in Figure 23 requires students to complete four different parts of an item:

18.

x	*y*
0	5
1	10
2	20
3	40
4	
5	

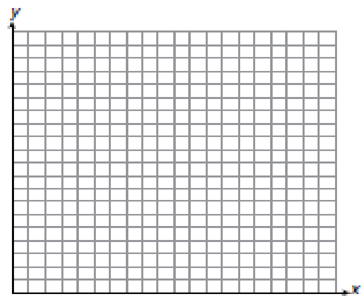
Part A Given the table, what are the next two entries?

____

____

Part B Write a rule in the space below that shows the relationship between the *x*-values and the *y*-values in the table.

Part C Graph the relationship from *Part A* and *Part B* on the grid. Be sure to mark your scale on your axes.



Part D What does the graph tell us about this relationship?

Figure 23. Item 18
(Reprinted with permission from the Denver Public Schools)

According to the curriculum specialist reviewing this item, the first three tasks correspond to concepts that many proficient and partially proficient grade 8 math students should be able to complete. However, the final task would likely be missed by both lower and higher performing students who may not have the foundation to properly describe the relationship as one of three possible answers given in the rubric: “not linear”, “exponential” or “quadratic”. The specialist hypothesized that fewer students have the opportunity to earn full points on CR item 18 since the last part of this item taps into an algebraic foundation outside of the regular math track. In other words, this item may not be differentiating between the performance of proficient and partially proficient students largely due to the final task that must be answered correctly in order to earn full points on this item.

The holistic rubric was also reviewed to determine whether the scoring rules may be contributing towards the finding that the item does not appear to clearly distinguish between the performance of proficient and partially proficient students. The holistic rubric below CR item 18 provides the following guidelines to teachers:

4 points: Responses include 4 correct answers and explanations and work that shows how the student arrived at the answers. (note for Part D: The relationship [on the graph] is “exponential” or “not linear” or “not quadratic”).

3 points: Responses include 3 correct answers and work shown for all problems demonstrating how the students arrived at the answers.

2 points: Responses include 2 correct answers and 2 incorrect answers.

1 point: Responses include 1 correct answer with work shown, or enough work shown to demonstrate some understanding.

0 points: Responses demonstrate no evidence that students has mathematical knowledge of how to represent, describe, and analyze patterns and relationship using tables, graphs, verbal rules, and standard algebraic notation.

In contrast to the rubric presented for item 12, the rubric for item 18 clearly distinguishes the separate conditions that students need to meet in order to land in each scoring category. A rater reviewing a response should be able to easily apply the rubrics since each category is contingent upon the exact number of correct responses to each part of the item and there are no other conditions to consider. For this item, it would seem less likely that the scoring rubric contributed to raters granting similar scores to proficient and partially proficient students, but rather that earning full points required exhibiting knowledge located outside of the grade 8 standard math track.

In summary, the findings from this chapter suggest that the scored CR items requiring rater judgment on the grade 8 math test may not be reflective of the true performance of students. As indicated from the findings, the evaluation of each item revealed that all three CR items were flagged by the point-biserials due to potential problems with how raters rated student responses to individual items, how students interpreted the prompt, or the inclusion of items that are located outside of the subset of skills and knowledge being measured by the interim test and the larger grade level math target domain. In the specific case of the grade 8 math test, items 6 and 18 tested content outside of the grade level math target domain, and the wording of item 12 could have been refined to ensure that more students had the opportunity to earn full points on this item. In addition, for item 12, the rubric guidelines should have reviewed more carefully to ensure that there is more distinction being made between the performance of students who are "partially proficient" and "proficient" in the two-point category. Although the issue raised by the curriculum specialist concerning the inclusion of content that is not reflective of the traditional curriculum or the pacing of the curriculum may not apply to all interim tests, the low point-biserials and reversed point-biserials found for eight out of nine tests in each level and for all three content areas suggest that these problematic items should be re-evaluated to determine whether individual items may need to be replaced or refined. Some of the steps undertaken to evaluate CR items flagged for having undesirable qualities reflect examples of the type of work that ideally should have taken place during a piloting phase. Further, because of the limited number of items included in each interim test, careful item selection should have been instituted to ensure that items for each interim assessment reflected content covered and learned by most students to ensure that each item provides good information about the majority of students taking these tests[18].

As mentioned in the beginning of this dissertation study in Chapter 1, the rapid process used to develop these interim assessments in this district exemplifies a process that has and is still taking place in many other school districts nationwide using the same testing company or developing their own set of interim assessments. The findings under this section point to issues that could have been addressed if a piloting and professional development period, such as the process described by Vendilinski et al. (2007), was built in to better check for item quality. As described earlier in the literature review, the researchers attributed the development of quality items meeting technical standards to the time given by the district to carefully co-develop these tests with teachers.

The first year of administering these tests in DPS represented a year where the properties of the tests were described by staff as “largely unknown”. Again, considering that many other districts, such as the case study districts referred to in the literature review, engaged in the same process of rapid deployment, the staff’s description of the tests administered throughout the district in 2006-07 as “largely unknown” would likely apply to tests administered in many other school districts during the first year of implementation. Unfortunately, within the context of this district and many other school districts, these

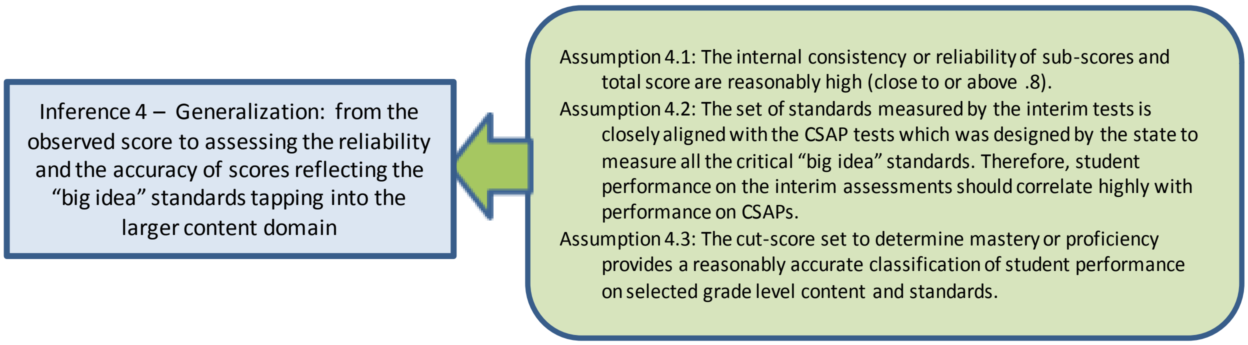
[18] The author observed an item panel session for a grade 10 geometry test being developed for the 2008-09 year. In developing the 202008-09 tests, DPS assessment staff included the point-biserials for item panels to review. In that session, panel members voted to keep an item with a negative point-biserial. According to one panel member, the rationale for keeping that item was that “…it would tell me if one or two of my foreign students are coming in with high math skills”.

tests were used during the first year of implementation to drive decisions despite the fact that all of the items being used were new and little time was made available during the school year to evaluate and modify each batch of items prior to administering them to students.

Although district staff substantially improved the process for evaluating each item during the second year of administering the interim tests and subsequent years, many new items were added to each 2007-08 interim test. In 2007-08, the item panels added new items to the math and reading tests and replaced more than thirty percent of old items used in 2006-07 with new ones furnished by TPR. Considering the limited timeframe for carefully evaluating each item, the addition of many new items to all interim tests would likely result in detecting similar findings to those presented in this chapter. That is, a few items that do not appear to add value in the assessment of students are flagged, after those items were used to drive different uses and decisions. In the case of this school district, although district staff believed that the test development process improved considerably from the first year of implementation, the rapid cycle of refining and administering these tests during the school year still posed significant challenges in the second year. For example, as noted by assessment staff members, “our math panels have improved tremendously, but we’re having a hard time trying to find a committed body of teachers who can commit to the literacy item panels…for some meetings, attendance is low with only three panel members showing up”.

In the next section, the evaluation of the interpretive argument moves away from evaluating individual items to determining whether the overall instrument appears to provide reliable information for making inferences about student performance over the subset of standards used to measure grade level skills and knowledge.

Inference 4: Generalization



This section focuses on evaluating the assumptions supporting the generalization inference. The three assumptions characterize how well the observed score for each student presents a reliable or precise representation of a student's performance over a subset of state standards. As discussed previously in Chapter 3, the target domain consists of the entire range of grade level skills and knowledge in each content area (state standards) and the subset represents a selected set of those skills and knowledge given to students within a formal testing context. In a test development context, the set of conditions used to evaluate the subset of the target domain represented on each test typically includes the type of tasks or items selected to evaluate a student's performance, the number of raters used to score student responses and the number of times or occasions that a test is given to assess students over a time period. In this section and as detailed earlier in Chapter 3, the evaluation of the Generalization Inference focused largely on examining the precision of scores earned from responding to a set of CR and MC items on the interim tests. The purpose of checking these assumptions under the generalization inference is to determine the extent to which these scores and proficiency judgments are representative of student performance over the subset of standards tested within a formal testing context. If precision appears to be relatively high, this finding would provide evidence supporting the use of these tests to support inferences based on proficiency over the content represented in each test. In this section, the analyses conducted provide evidence addressing the extent to which interim assessments meet the reliability standard of .8 used by the

CSAP and by many other test developers, whether sub-scores assessed on the grade 8 math test appear to provide reliable information about students, and how well the scores and proficiency classifications earned by students correspond to the CSAP scores and proficiency classifications earned by the same group of students.

The first assumption checked under the generalization inference is the reliability of each of the nine assessments. To evaluate the extent to which the test provides reliable information about student performance on the set of items chosen to assess grade level standards, the reliability of each instrument was calculated using Cronbach's Alpha. Table 10 shows the alpha level for each of the nine tests reviewed. As seen in Table 10, six out of nine tests displayed reliabilities above the standard threshold of .8, and the other three tests had values that were very close.

Table 10

Cronbach's Alpha for Selected Interim Tests

Grades	Content Area	Version	Years	Alpha Levels	Number of Items
4	Math	1	2007-08	0.81	23
4	Math	3	2007-08	0.85	23
4	Reading	1	2007-08	0.86	25
4	Reading	3	2007-08	0.87	25
8	Math	2	2006-07	0.75	18
8	Writing	2	2006-07	0.85	19
8	Writing	2	2007-08	0.79	19
10	Reading	1	2006-07	0.84	26
10	Math	2	2006-07	0.77	17

Based largely on the alpha levels presented above, it would seem that most instruments can reliably measure students of varying proficiency. These findings suggest that despite identifying a few items under the scoring inference with undesirable qualities (e.g., low point-biserials) on eight out of nine interim tests, these items do not have a considerable effect on the instruments' overall reliability. That is, most of these tests can provide reliable or consistent information to differentiate the performance of low and high achieving students. The three tests that did not meet the .8 threshold represent tests that consist of fewer than 20 items. Since the number of items on a test directly influences the strength of the reliability coefficient, increasing the number of test items could serve as one strategy for strengthening the reliability coefficient for those tests. However, the earlier finding of local dependence also needs to be considered when evaluating the reliability of each interim test. That is, the reliability estimates presented not just for grade 8 math, but for other tests are likely biased upwards due to presence of local dependence affecting each test. Although it is difficult to know the degree to which local dependence is inflating the reliability for each test, in the case of the grade 8 math test, the effect is probably considerable due to the high degree of local dependence detected across all pairs of items. In addition to assessing the reliability of each total score, since proficiency classifications for each student are established relative to each standard reported on the test, the reliability associated with each subscore is also evaluated here. As done for the previous sections, the next set of analyses focus largely on testing out the assumptions under the generalization inference using the grade 8 math test.

During the two school years evaluated for this study, scoring reports provided to teachers provided proficiency judgments about students based on the overall score, the Colorado standards represented, and the CSAP framework statements. C-5 in Appendix C shows an example of a student level scoring report for a grade 7 student taking the version 3 math test in the 2007-08 year.

The sample student scoring report in Appendix C-5 reflects the same scoring report that a teacher would see for an individual student in any other content area and grade. As seen in Appendix C-5, the individual score report provided teachers with the student's overall proficiency score, and noted whether the student was "at or above proficient" or "below proficient" on each standard and each framework statement. According to DPS assessment staff, the subscores were provided to teachers on scoring reports to help teachers identify individual student weaknesses and strengths on state standards being tested for each content area. Teachers were then encouraged to take appropriate interventions and actions based on the information provided from the subscores. The scoring reports were also produced at aggregated levels (e.g., by classrooms, by grade and by school) and the aggregated reports were also used to drive decisions at those levels. For example, one teacher interviewed for this project noted that the teachers of his English Department implemented the practice of re-teaching or focusing on a specific content standard where low performance consistently surfaced across scoring reports.

In addition to utilizing scoring reports to evaluate individual student and classroom needs, some teachers during the 2006-07 and 2007-08 years evaluated student growth using the subscores at the standards level for Procomp merit pay. Although the practice of setting Student Growth Objectives (SGOs) based on performance by standards is currently discouraged by DPS administrators, ProComp staff noted that some teachers developed student growth objectives using the percentage of student scoring proficient on selected standards during the 2006-07 and 2007-08 years .

Herman and Baker (2005) and Briggs and Wilson (2003) note that information from subscores can provide useful information for teachers to determine how well students understand different strands of a content area. However, these researchers also point out the importance of ensuring that the subscores reported should be reasonably reliable if useful information is to be gained from understanding performance relative to each dimension assessed. The grade 8 math test, like all math tests administered in 2006-07 and 2007-08 had a total of six standards. For this grade 8 math test, a student, classroom or school received a proficiency rating on each of the six standards represented on the test.

Standard 1 – Students develop number sense and use numbers and number relationships in problem solving situations and communicate the reasoning used in solving these problems.
Standard 2 – Students use algebraic methods to explore, model, and describe patterns and functions involving numbers, shapes, data, and graphs in problem-solving situations and communicate the reasoning used in solving these problems.
Standard 3 – Students use data collection and analysis, statistics, probability in problem-solving situations and communicate the reasoning and processes used in solving these problems.
Standard 4 – Students use geometric concepts, properties, and relationships in problem-solving situations and communicate the reasoning used in solving these problems.
Standard 5 – Students use a variety of tools and techniques to measure, apply the results in problem-solving situations and communicate the reasoning involved in solving these problems.
Standard 6 – Students link concepts and procedures as they develop and use computational techniques, including estimation, mental arithmetic, paper-and-pencil, calculators, and computers, in problem-solving situations and communicate the reasoning involved in solving these problems.

Figure 24. Standards represented on grade 8, version 2 math 2007 test and all middle school interim math tests administered in 2006-07 and 2007-08

To evaluate the reliability of subscores, separate reliability coefficients were generated for each standard using Cronbach's Alpha.

Table 11 displays the reliabilities associated with each standard where separate subscores and proficiency levels were reported to teachers.

Table 11

Consecutive Reliabilities for Grade 8 Math Test

Reliabilities by Standard					
Standard 1	Standard 2	Standard 3	Standard 4	Standard5	Standard 6
n/a	.6	.35	n/a	n/a	n/a

As seen in Table 11, the highest reliability obtained for all six of the standards assessed was .6 for standard 2. Standard 3 has a reliability of .35. The reliability coefficient is not reported for standards 1, 4, 5 and 6, since those dimensions are assessed by only one item. Although the analysis of evaluating subscore reliability was only conducted using one test, similar findings would likely apply to the other eight interim tests. In 2006-07, all math tests comprised of 18 items, measured at least five standards, and reported proficiency judgments on each standard. In 2007-08, five more items were added to every math test, with a minimum of five standards reported for each test. Writing comprised fewer items with a total of 18 items available on each test for both the 2007-2007 and 2007-08 tests. The district reported subscores for a minimum of five standards on each writing test. For the reading assessments, although there were more items populating those assessments (an average of 25 items across all grades and for 2006-07 and 2007-08 school years), the same concerns apply since each test had at least one standard that was evaluated with fewer than four items. Considering that subscore reliability may be low (below .5) for several standards assessed by each test, the scores for at least one standard on each interim test may not be considered reliable enough to support the current practice of encouraging teachers to make instructional decisions for individual students using the subscore information. Based on the levels of reliability found for each subscore on this test, these findings may suggest that DPS may want to avoid reporting proficiency judgments for each standard and simply report the proficiency level attained by an individual student based on overall performance.

At the classroom level, and for the purpose of supporting the past practice of setting performance objectives based on subscore performance, the reliability based on decisions made at aggregated levels relative to individual students may differ. That is, the scores for groups of students would theoretically be estimated with more precision than individual scores, since the error associated with the group mean is typically smaller than the error associated with an individual's score[19]. Therefore, even if standard 2 has a reliability of approximately .5, this may suffice for driving actions impacting groups of students. This issue is discussed further in the next set of findings.

The SEM discussed earlier under the item design inference for the same grade 8 test is revisited in this chapter. Under the item design inference, the conditional SEM generated using a PCM was evaluated to determine whether there were sufficient items on the grade 8 math test to assess students of varying proficiency. The findings pertaining to that study also address the generalization inference, since the previous study's objective of determining whether a test is populated with sufficient items also serves as a good indication of whether those items can precisely evaluate students at varying proficiency. As found previously in the item design inference section of this chapter, although the SEM was found to be relatively flat across varying levels of proficiency, the SEM was also found to be considerably large at .4 across most respondents. The implications of this finding were discussed briefly in that section and will be discussed again here. Figure 25 presents the SEM plot shown earlier under the item design inference and shows that the lowest point of the SEM is located at approximately .4.

[19] However, Brennan (1995) argues that the assumption that the SEM would be lower for groups is rarely disputed. His paper cautions against not evaluating this assumption and provides cases where this assumption does not hold.

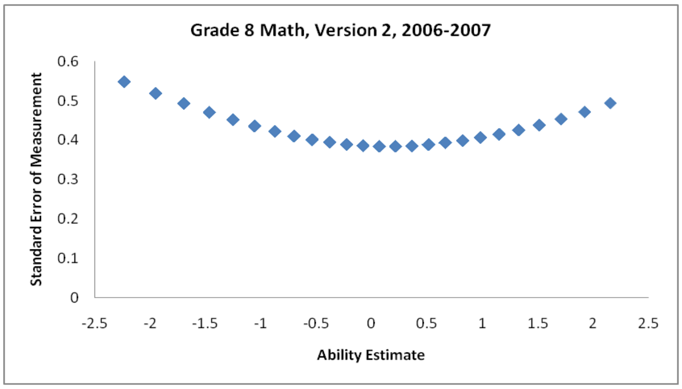


Figure 25. SEM Plot for Grade 8 Math Test

Considering that the SEM is relatively flat, the implications of this finding is that the SEM can also be evaluated using a simpler classical test theory (CTT) approach which assumes by convention, that a constant SEM estimate can be applied across all respondents. Evaluating the SEM based on the raw scores under CTT would also directly correspond to the district's use of raw scores to make proficiency determinations. Under CTT, the standard error of measurement for individual scores is calculated using Equation (4.6).

$$\sqrt{1-r} * \sigma \tag{4.6}$$

Factoring in a test reliability of .75 and a standard deviation of 4.6 on this test, the SEM is equal to 2.3. If constructing a .95 confidence interval, a student's performance could be located between approximately plus or minus 4 points from the student's score.

To evaluate the precision of student scores, two confidence intervals were applied to assess the error estimate around each student's score: a more stringent .95 and a relaxed .67 confidence interval. The application of these two intervals provided information on whether absolute decisions or strong generalizations about student mastery over the subset of content being measured by this test may be supported. Under the item design inference, the less stringent confidence interval was applied to hypothetical students located at the unsatisfactory and partially proficient locations. Based on a less stringent confidence interval associated with scores for these two individuals, it appeared that almost all students in the partially proficient region could be found in either the proficient or unsatisfactory region, and 376 unsatisfactory students could be located in the partially proficient region.

Under a CTT approach, the implications of factoring in a confidence interval around each student's total score at the .67 level are reflected in the crosstabs depicted in Tables 12 and 13.

Table 12

Comparing Performance Using a .67 CI

	Proficiency At Lower Bound			
Original Classification	A	P	PP	U
A	2	17	0	0
P	0	54	122	0
PP	0	0	113	291
U	0	0	0	N/A

Table 13

Comparing Performance Using a .67 CI

	Proficiency At Upper Bound			
Original Classification	A	P	PP	U
A	N/A	0	0	0
P	54	122	0	0
PP	0	247	157	0
U	0	0	376	3024

In Tables 12 and 13, the highlighted cells represent the total number of students who would be classified in a different performance band when factoring in the SEM using a .67 confidence interval. Based on the numbers reflected in the highlighted cells for both tables, approximately a third of all students could potentially fall into different performance bands when applying a less stringent confidence interval around each student's observed score. At the more conservative .95 level, the total number of students who could fall into more different performance bands would increase considerably. Tables 14 and 15 present performance outcomes based on using the more conservative interval of .95.

Table 14

Comparing Performance Using a .95 CI

	Proficiency At Upper Bound			
Original Classification	A	P	PP	U
A	N/A	0	0	0
P	176	0	0	0
PP	0	404	0	0
U	0	203	656	2541

Table 15

Comparing Performance Using a .95 CI

	Proficiency At Lower Bound			
Original Classification	A	P	PP	U
A	0	19	0	0
P	0	102	74	0
PP	0	0	0	404
U	0	0	0	N/A

The application of a higher confidence interval to the total scores indicates that all students in three performance bands (advanced, proficient and partially proficient) could be estimated in different performance bands. The distribution of students in each proficiency level when applying a .95 confidence interval is similar to the distribution found earlier under the item design inference when a conditional SEM was explored using a .67 interval. That is, all of the partially proficient students at the .95 level could be classified in other proficiency bands.

When using the interim tests for diagnostic or classroom purposes, the consideration of the SEM relative to a confidence interval may be unnecessary since a teacher could ideally draw upon other points of data to better inform that teacher’s contextual framework for understanding what the student knows. However, when attaching higher stakes to this test, in particular when a single assessment system is used to drive decisions and resources such as identifying students for mandatory summer remediation or paying teachers on the basis of how many more students became proficient between one interim test to another, the confidence intervals provide a useful indication of whether those decisions were supported by precise information from the test.

Based on the considerable percentage of students who could fall into more than one proficiency category in Tables 12-15, these findings raise questions about the soundness of driving decisions affecting individual students based solely on using the interim assessments. In addition, since the findings under the item design inference revealed that the size of the SEM on this grade 8 math test was also found to be similar to the SEM found across all other nine tests reflecting different grades and content areas, this finding suggests that the different uses of these interim tests may need to be considered in conjunction with other evidence to support decisions attached to individual student scores.

As noted earlier, for uses pertaining to aggregated levels, one would by convention, expect the standard error to be lower based on the number of students factored into the equation. For example, if evaluating the entire population of test-takers on the grade 8 math test, the error associated with the group mean of all total scores can be calculated by dividing the standard deviation of group scores by the square root of the total number of students taking the test. For the population of grade 8 students (n=3,999) and a standard deviation of 4.56, the standard error of the mean is equal to approximately .07. At the classroom level, assuming that the standard deviation is equal to the population's standard deviation of 4.56 and based on an average classroom size of 25 students, the standard error of the mean is equal to approximately .77. The considerably smaller standard error found at aggregated levels provides some supportive evidence that it may be reasonable to utilize these tests to support decisions based on aggregated or grouped scores. However, finding a smaller standard error of the mean does not resolve the qualitative issues related to just the CR items evaluated in the previous scoring inference section.

The last assumption checked under the generalization inference evaluates the extent to which the interim assessments and the CSAP provide similar performance information about students. As mentioned in Chapter 3 because each interim assessment is designed to evaluate students over the same set of standards weighted more heavily in the CSAP tests, a student's proficiency on the interim test should logically align with the proficiency earned on the broader CSAP tests. To evaluate this predictive relationship between the interim tests and the CSAP tests, the total CSAP scale scores were correlated against the total raw scores earned by the same group of students who took the interim assessments. A moderate to strong correlation found between the two test systems may suggest that student performance on the interim assessments could provide insight into expected performance on the CSAP.

Table 16 displays findings from correlating the total scores earned on the interim assessments with the total scale scores earned by students on the CSAP. The median correlation in the table equals .72 and the correlations range from .67 to .84. The findings from Table 16 suggest that student performance on the interim assessments are strongly associated with performance on the CSAP. Although one would expect performance to be less strongly correlated on the version 1 tests, the table suggests that teachers may expect that students who performed either poorly or highly on versions 1, 2 or 3 of a test may perform in a similar manner on the state test system.

Table 16

Correlations between Interim Tests and CSAP

Grades	Content Area	Version	Years	R	Number of Items
4	Math	1	2007-08	0.72	23
4	Reading	1	2007-08	0.7	25
4	Math	3	2007-08	0.84	23
4	Reading	3	2007-08	0.77	25
8	Math	2	2006-07	0.72	18
8	Writing	2	2006-07	0.72	18
8	Writing	2	2007-08	0.67	18
10	Reading	1	2006-07	0.71	26
10	Math	2	2006-07	0.73	17

Table 16 also indicates that the correlations between tests strengthened in the 2007-08 year and with each version 3 test. Although the correlations between each version 1 and 2 test and the CSAP are strong, the stronger association found between the CSAP and the version 3 tests suggest that the assessment staff's objective of ensuring that the version 3 test represents summative content similar to the CSAP appears to have been met. Furthermore, considering the extent to which stakeholders believe that the CSAP represent a good measure of proficiency on state standards, this finding would also suggest that these interim assessments may provide good information about student proficiency before the CSAP results are released. This finding that the interim assessments are predictive of how students will perform (if viewing versions 1 or 2 of each administration) or have performed (if viewing results from version 3 of each administration) may also help mitigate uncertainty or surprise to central administrators when test results are released to all school districts by the end of each summer.

Having seen that all versions of the interim tests may have helped provide district staff with insights into how students performed on the CSAP tests, the next analysis evaluates the extent to which both test systems attribute similar proficiency classifications to students. For this analysis, the proficiency classification of students taking the grade 8 math test was compared relative to the proficiency classification of the same students (n = 3,753) who took the grade 8 math CSAP test a few months later. The comparison entailed using a cross-tab to examine student classifications on the grade 8 math CSAP relative to the grade 8 math interim test. The cross-tab of proficiencies presented in Table 17 shows that in 2006-07, 924 students or 24% of the total grade 8 population earned a rating of “unsatisfactory” on the grade 8 math test, but received a higher performance classification on the corresponding CSAP test. The proportion of students that may have been misclassified as “unsatisfactory” by the interim test is similar to the proportion found earlier when evaluating the constant SEM using a .95 confidence interval. In that study, the findings revealed that a similar number of “unsatisfactory” students (859 students) could have been classified as either “proficient” or “partially proficient”. Further, both the sets of findings indicate that approximately one third of all grade 8 “unsatisfactory” students were potentially misclassified as candidates for remediation.

Table 17

Comparison of performance classifications between grade 8 Math CSAP and grade 8 Math

	Proficiency on CSAP Math				
Proficiency on Math Interim Assessment	A	P	PP	U	Total
Advanced	17	2	0	0	19
Proficient	147	209	55	2	413
Partially Proficient	74	301	296	67	738
Unsatisfactory	20	176	797	1590	2583

Table 17 also highlights notable discrepancies in classifying students at the proficient and above levels. As seen in the table, the interim test identified 275 fewer students as “proficient” relative to the CSAP and only identified 19 as “advanced” compared to 258 identified as “advanced” for the CSAP. As indicated earlier in this chapter, since data from the version 2 tests were used by some teachers during the first two years of implementation to re-set their student growth objectives, teachers who re-set their objectives based on this grade 8 math test may have been compelled to lower their objectives based on the considerably small proportion of students who were classified as proficient on this one test.

As described earlier in Chapter 2, during the 2007-08 year, DPS established a new process for setting cut-points for the grades 3 through 8 interim assessments where proficiency cuts were allowed to vary for each content area. Staff contended that this new process yielded proficiency judgments that appeared to be more aligned with how teachers would generally classify their student’s mastery over content (DPS Assessment Staff, 2008). However, although the cuts varied by content area, the same cuts were applied across all elementary and middle school grade levels. This method was also applied to the course assessments not evaluated in this dissertation study, but as reported by the media in 2007, one high school typically rated each year as an “excellent” or “high” performing high school in the entire state, only had .8% of the total population of students receive an advanced rating on the interim tests. Although these performance misclassifications can occur when the cut-points or standards for proficiency may be set too high or too low for a given assessment (see for example, “Educational Achievement Standards: NAGB’s Approach Yields Misleading Interpretations”, US General Accounting Office, 1993), the findings under the item design inference and in this section suggest that it may also not be reasonable to differentiate four performance categories based on the small number of items available on these interim tests. Each interim test comprises at the most, a third of the minimum number of items present on a given CSAP test. Considering that the findings under the item design and generalization inferences suggest that there appears to be an inadequate number of items to differentiate between different proficiency levels, it seems unlikely that there are adequate items to distinguish performance differences between four proficiency categories.

Summary

Assessed together, the findings throughout this chapter suggest that the practice of using these interim tests for making absolute decisions affecting individual students does not appear to be strongly supported. The evaluation of the interpretive argument in this chapter consisted of checking the assumptions specified under the item design, the scoring, and the generalization inferences. Under the item design inference, although the findings supported the assumption that each of the nine interim tests examined was populated with items that students of varying proficiency could answer correctly, the findings under that section indicated that each test had an inadequate number of items to clearly establish the proficiency estimate and classification of many students.

Under the scoring inference, finding MC items with low-item total correlations (below .3) on each of the nine tests and flagging CR items on five out of nine tests with potentially problematic qualities indicate that some of the items could have been better refined for the purpose of acquiring more meaningful or better information about student performance. Further, the findings under the scoring inference point to the difficulty of differentiating the performance of "proficient and above" students from "below proficient" students as a result of either unclear scoring guidelines provided to raters, the wording in the item, or the inclusion of content that was outside the grade 8 math target domain. As noted earlier in this section, the inclusion of items measuring high school algebra content was continued in the 2007-08 year to serve the purpose of determining which students could skip Algebra I in grade 9. Although the inclusion of advanced items is useful for students with higher math abilities, considering that these test are not offered in a computer adaptive format and that that there is a limited number of items on these math tests and (18 in 2006-07 and 23 in 2007-08), the interim assessments may not serve as the best place for obtaining information about determining algebra proficiency. In particular, since the reliability of information pertaining to algebra proficiency would not be adequate based on data provided by three items on the test.

The final set of findings presented under the generalization inference, reinforces evidence illuminated earlier under the item design inference section. That is, when factoring in the SEM at either the .67 or .95 levels, a considerable number of students could belong to more than one performance category. As shared earlier under the item design inference, the size of the SEM was similar to the SEM found for the grade 8 math test on other interim assessments from the 2006-07 and 2007-08 years and this would suggest that students were assessed with similar levels of precision as the grade 8 math test.

In addition to the findings associated with the SEM, the comparison of proficiency levels on the CSAP and interim tests suggest that the uniform cut-scores established in 2006-07 for all interim tests and grades classified more students as “below proficient” on the interim tests relative to the CSAP. Since the scores on all nine of the interim tests reviewed are moderately to highly correlated with the CSAP, this finding would indicate that the interim tests may have fulfilled one purpose of being used as predictive instruments for the CSAP. However, since all decisions are driven by the proficiency levels rather than the scores, the performance discrepancies found between the two test systems for the grade 8 math test revealed that the cuts on the interim tests may have been set too high for the purpose of driving higher stakes decisions, such as mandatory remediation, in 2006-07. As suggested earlier under the item design inference, one strategy for classifying students would be to only differentiate the performance between those who are below proficient from those who are at or above proficient. This simplified performance structure would help address the problem associated with having to better differentiate four distinct categories using both limited items and narrow score ranges.

The evidence from evaluating the item design, scoring, and generalization inferences in this chapter, appear to challenge the use of the grade 8 math test in 2006-07 for identifying individual students for mandatory remediation in the summer and for predictive assumptions about the test such as leveling students by ability. However, in this particular case study and within the context of so many other districts (such as the two case study districts reviewed in Chapter 3) utilizing the same approach of co-creating interim tests with a test vendor, non-piloted tests were used to inform multiple decisions. Again, although these tests may have provided helpful information in certain low-stakes contexts (e.g., providing extra tutoring services to lower performing students), the three inferences of the interpretive argument evaluated in this chapter suggest that the test may not have provided adequately reliable information for driving decisions that have higher consequences for individuals such as mandatory remediation or assessing the effectiveness of teachers to improve student performance.

Although the evidence presented in this chapter may challenge the uses of the tests at the individual student level, adequate evidence was not collected to assess the uses of the test at aggregated levels. The next chapter evaluates the arguments supporting the decision to use these interim assessments as measures for teacher effectiveness in improving student performance. In contrast to the sets of studies conducted in this chapter that evaluated the extent to which the interim test data supported the uses of the tests for making decisions impacting individual students, the next chapter focuses predominantly on the use of individual scores aggregated to the classroom level.

CHAPTER 5 – Evaluating the Use of Interim Assessments for Merit Pay

As indicated at the end of the previous chapter, this chapter focuses on evaluating the assumptions supporting use of the interim assessments to evaluate and reward teachers. The findings from evaluating Inferences 2, 3 and 4 in the previous chapter have direct implications for using these interim tests for teacher merit pay (Use 3) since the percentage of individual students moving from lower to higher performance bands between test versions are typically used to evaluate and reward teachers in this district. A key and consistent finding that surfaced in the previous chapter was that evidence collected in the different analyses was that there was considerable uncertainty associated with making absolute decisions affecting individual students based solely on the interim assessment data. Although the findings from the inferences evaluated earlier suggest potential problems with using the individual counts of students in each performance band to establish pre- and post- test performance classroom performance for merit pay, the primary purpose of this chapter is to establish whether these tests can serve as valid measures of student growth.

In the previous chapter, the grade 8 math, version 2, 2006-07 test was used to test out the assumptions under the item design, scoring and generalization inferences. This test was selected since it met both of the uses being evaluated in that chapter: identifying individual grade 8 students for mandatory remediation during the summer of 2007 and for identifying individual students who may need additional remediation or tutoring prior to taking the CSAP tests. In this chapter, math and reading tests from a different grade (grade 4 version 1 and version 3) and school year (2007-08), were used to test out the assumption under Use 3. The studies in this chapter evaluated different interim test versions (version 1 and 3 used within the same year) since teachers set SGOs using data from the versions 1 and 3 tests.

The decision to use the grade 4 interim tests, as opposed to other tests available in the middle and high school grades, was based largely on being able to easily match students in grade 4 to teachers having instructional responsibility over those students in math and literacy. Since district staff at the time could not verify primary instructional responsibility for students with duplicate assignments to core subject teachers in the middle and high school grades, the studies in this chapter used the grade 4 tests where most teachers instruct all core subject areas. In addition, the 2007-08 tests were used since a new method for setting cut-points described earlier in Chapter 2 were developed to assign students to a performance band beginning in the 2007-08 year to the present.

This chapter begins with a brief overview of ProComp to situate the use of the interim tests for merit pay purposes. Following the overview of ProComp, the subsequent section briefly discusses the approach outlined in Chapter 2 and used by DPS staff to assess student growth across test versions. After outlining the approach used by DPS to evaluate student growth between tests, an overview of the vertical scaling approach commonly used by test developers to create compare scores across tests administered to students of differing proficiency is presented. This overview of vertical scaling provides context for understanding the approach used in this study to place the scores of the interim onto a common and comparable scale. This chapter culminates with presenting the findings from testing out the assumptions supporting the use of these interim tests to evaluate and reward teachers based on student growth.

Overview of ProComp

Under ProComp, teachers are evaluated and compensated through four different components: knowledge and skills; comprehensive professional evaluation; market incentives; and student growth. The ProComp chart located in D-1 of Appendix D presents the four main components of the ProComp system back in 2007-08. Although the ProComp system has evolved considerably over time to include different elements for teacher compensation, the student growth objective element described in 2007-08 remains intact. In 2007-08, the Student Growth Objective element in Appendix D-1 represents the only area where all teachers are given the opportunity to earn either a bonus or a salary increase based on setting ambitious and attainable growth objectives. All other achievement-related incentives require teachers to exceed goals relative to CSAP exam scores.

As described earlier in Chapter 3, every teacher is expected to set two different student growth objectives (SGOs) using students continuously enrolled in the classroom since October 1 through the end of the academic school year. All teachers set two objectives based on data from two different types of assessments (e.g., one math interim assessment and one reading interim assessment, or one writing interim assessment and the Direct Reading Assessment). Teachers set their student growth objectives using the version 1 test results as baseline information about students. Both teachers and principals use the version 3 test data to evaluate whether the objectives specified in the beginning of the school year were met. A merit pay bonus was awarded to teachers who met one SGO. Teachers who met both objectives received a merit pay bonus and a salary increase. For example, a new grade 5 schoolteacher who joined the district in 2007-08 with a class of 30 students may have set two SGOs as follows:

1. Out of the 20 students who scored “unsatisfactory” on the version 1 math interim assessment, 15 out of 20 or 75% of those students will move out of “unsatisfactory” on the version 3 math interim test.

2. Out of the 10 students who scored “below grade level” on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) test, 8 out of 10 students or 80% of those students will be considered as reading at “grade level” by the end of the year.

Based on information in the chart presented in Appendix D-1, if this teacher meets the objective set for the interim test but fails to meet the objective set for the DIBBELs test, the teacher would receive a merit pay bonus of $356. If the same teacher meets both objectives, the teacher would receive the bonus of $356 and a salary pay increase of $356 (each of these reflect 1% of the base index set by ProComp in 2007-08 of $35,568) or a total of $712 for the year.

According to ProComp staff, in the 2006-07 and 2007-08 school years, a large number of teachers in the district created SGOs using two different interim tests for evaluating student growth or learning taking place in their classrooms. Teachers were encouraged to use the interim tests by instructional leaders and district administrators since these tests represented one of the only instruments available at the time, which presented data common to all classrooms and schools, and mapped directly to the CSAP frameworks. According to ProComp staff, two SGOs commonly developed by teachers using the interim test data specified the increase in the percentage of students scoring “proficient and above” on the version 3 test and /or the percentage of students moving up by one proficiency category[20]. In addition to evaluating whether the interim tests can serve as valid measures of student growth, this chapter also evaluated to what extent classroom outcomes evaluated by commonly defined SGOs can differ if using different scores adjusted for the difficulty found across interim test versions.

The next section provides background information on the steps taken by DPS staff in 2007-08 to compare student performance across interim test version. This background information is followed by an overview of "vertical scaling" or a standard statistical procedure used by test developers to compare test scores across different levels (e.g., grades). This overview of vertical scaling provides context for understanding the approach used in this study to place the scores of the interim test versions onto a common scale.

Background: Creating Comparable "Performance"

To understand the rationale for devising an alternative approach in this study for comparing student performance across interim test versions, some background information is provided on the standard-setting approach used by DPS to compare student performance across test versions. As shown by the p-value distribution referenced earlier in Chapter 4 under the Item Design Inference (see Appendix C-2), the third version of each test was intentionally designed to be more difficult than the first version of the test administered at the beginning of the school year. In 2006-07, the same set of performance cuts (see Chapter 2 for details on cuts) tests were applied across all three test versions, grades and content areas. These cuts were selected as a best guess estimate of how students at varying levels of proficiency would likely perform on any given interim test.

As described earlier in Chapter 2, in the following year, district staff recognized that the performance bands needed to be adjusted to account for differences in difficulty found across test versions. According to a document released by the assessment department in 2007-08, "the proficient bands were adjusted for each test to allow for the comparability of proficiency performance between the

[20] See for example, the Student Growth Objectives Handbook available at: http://denverprocomp.dpsk12.org/support/sgo

first and third versions of each test". Adjusting the performance bands entailed having the item panels evaluate each individual item on the high likelihood of being answered correctly by a student classified in one of four proficiency categories and selecting or modifying items of varying difficulty to fit the new cut points established for the 2007-08 tests. Since the versions 2 and 3 tests were designed to be more difficult than the version 1 tests, the number of points needed to reach proficiency on those tests was set at a lower level than the number of points needed to be proficient on the version 1 tests.

The performance band adjustments undertaken in the 2007-08 school year represented a substantial improvement from the previous year where the same cuts were applied across all three versions of the test, grades and by content area. In contrast to the 2006-07 school year, item panel members spent considerable time evaluating each individual item and selecting items that conformed to different proficiency profiles of students. However, despite the new approach takento use cuts that reflect differences in difficulty between tests by subject and version, all items had equal scoring weight regardless of difficulty and there was no evidence collected to verify how well each item corresponded to the assigned performance band.

In terms of weighting all items equally, if item panels judged one MC item as likely to be answered correctly by an "unsatisfactory" student and another MC item as likely to be answered correctly by an "advanced" student, both of these items were given an equal scoring weight of 1. Since the point values do not differentiate between items with varying difficulty, it may be possible for two students to share the same total score but to have different response patterns suggesting that one student may have a higher level of mastery over content assessed than the second student. Therefore, even if the performance bands were adjusted, the performance band ranges may not accurately reflect the proficiency performance of students since the scores across both versions carry the same scoring weight.

In reference to the mapping of items selected to the performance cuts established, since there is no evidence documenting or verifying the extent to which the panel's judgments on each item provided reasonably accurate indications of proficiency expectations, the possibility remains that the some items selected may have been classified incorrectly. For example, a panel may come to consensus that partially proficient students could answer a set of MC items correctly. However, it may be the case that those items were more difficult than the panel members realized during the compressed time allotted to develop and create each interim assessment. If those items reflected content that only highly proficient students had a higher likelihood of answering correctly, classifying those items as "partially proficient" would distort the number of points attributed to the partially proficient performance band range. Again, although the process of evaluating the individual items represented a considerable improvement from the first year of implementing these interim tests, it remains unclear whether this standard-setting and item selection approach sufficed "to allow for the comparability of proficiency performance between the first and third versions of each test".

Vertical Scaling and Placing the Interim Tests on the CSAP Vertical Scale

In practice, when the performance of students at varying levels (e.g., different grades or at different time points of a school year) must be compared across different tests measuring the same construct, test developers employ a test score linking procedure referred to as "vertical scaling". Linking test scores through vertical scaling is most commonly accomplished using IRT approaches (Kolen & Brennan, 2004). Any IRT-based linking approach requires that a set of common items is shared between the different tests since data generated from those common items are used to link scores across tests. An example of a vertically scaled test is the CSAP tests referenced through this dissertation.

The CSAP tests were vertically scaled by the test developer, CTB, using the common-items nonequivalent groups design depicted in Figure 26.

.

Figure 26. A common-items non-equivalent groups design.
(from Kolen and Brennan, 2004)

Under this design, the two groups taking the different test levels differ in ability or proficiency and only one form is administered to each group during a given test date. For example, Form X in Figure 26 could represent the grade 3 CSAP reading test and Form Y could represent the grade 4 CSAP reading test. As shown in Figure 26, a set of common items are shared between the two tests. These common items, also termed as "anchor items", are embedded in test forms across adjacent grade levels.

Before linking the test scores based on information from the common items, the tests are first scale using specified IRT models. CTB scaled their tests using two IRT models: the 3 Parameter-Logistic (3PLM: Birnbaum, 1968) and the Generalized Partial Credit Model (GPCM: Muraki, 1992). Both the 3PLM and the GPCM are extensions of the Rasch and the PCM models described earlier. The 3PLM adds a discrimination parameter and a guessing parameter to the Rasch model described in Chapter 4. If the guessing parameter is set to zero and discrimination is set to a constant value (e.g. all items share the same slope or steepness in the ICCs described in Chapter 4), the 3PLM would become the Rasch model. The GPCM, unlike the PCM, allows the discrimination parameter or the steepness of the ICCs for each item to vary. If the discrimination parameter was set to a constant, the GPCM would take on the same form of the PCM. Although CTB uses the 3PLM and the GPCM, other test developers have used the Rasch and the PCM to establish vertical scales. A number of factors are considered when determining which IRT model to use. One commonly cited reason in the field for using more complex IRT models is that these models fit the data better and can generate more precise proficiency estimates than the Rasch family of models (Briggs & Weeks, 2009).

Regardless of which set of IRT models are used, the IRT property which facilitates linking across scores on different tests is "parameter invariance". Under vertical scaling, if the IRT assumptions of local independence and unidimensionality are met, and the specified IRT model fits the data, then all IRT models share the property of parameter invariance. The general concept behind parameter invariance is that the item parameters of an item should not change regardless of the characteristics of the respondent group, and that the proficiency estimate of a respondent should stay the same regardless of which test is taken. Therefore, if the same set of common items on the CSAP are administered to both grade 3 and grade 4 students, then under the invariance property, the items parameters generated should not differ between groups.

In practice however, if separate runs are conducted for each grade using the common items, the item parameters would likely differ for grade 4 students than for grade 3 students. The item parameters would likely differ between groups since the scales are fixed or anchored to specified values (e.g., 0 and 1) to give meaningful interpretations to a proficiency estimate. As described earlier in Chapter 4, a proficiency estimate generated under any IRT model is given meaning relative to a scale anchored to the distribution of either items or respondents. In the case of the CSAP, since the scale is anchored to respondents, zero represents the proficiency mean and one represents the standard deviation relative to each unique group assessed. The goal of linking under IRT is to adjust or transform the mean and standard deviations on a test or across tests so that the common item parameters generated for both grades are the same and the scores are placed onto a common scale. The linear transformation of $\theta^* = A\theta + B$ is used where *A* and *B* represent the linking constants used to adjust the means and standard deviation of the scale and θ^* represents the transformed score. Once an optimal set of linking constants are found, a linear transformation is performed to convert the item parameters of a given test or tests relative to a base test (e.g., the test chosen to compare the test scores against). Most linking procedures utilize the below set of transformations:

$$a_{(iY)} = \frac{a_{(X)i}}{A} \tag{5.7}$$

$$b_{(iY)} = Ab_{(X)i} + B \qquad (5.8)$$

$$c_{(iY)} = c_{(X)i} \qquad (5.9)$$

In all equations, the subscript (iY) refers to a given item on the base test referenced here as Y and (x)i represents a given item i on the test being transformed. *A* in Equations (5.7) and (5.8) represent the linking constant used to adjust the discrimination parameter of all items and *B* represents the linking constant used to transform the difficulty parameter. Equation (5.7) represents the transformation made to the discrimination parameter. Equation (5.8) represents the transformation made to the difficulty parameter. Under a 3-PL, the transformations only apply to the *a* and *b* parameters and therefore no transformation is reflected in Equation (5.9).

To illustrate, if the linking constants are known and A = .8 and B = .05 and the item parameters for MC Item 1 on Test the grade 4 test are: a = 2, b = 1 and c = .15, then Equation (5.7) can be applied to transform the discrimination parameter as follows:

$$= \frac{a_{(X)1}}{A} = \frac{2}{.8} = 2.5 \qquad (5.10)$$

Equation (5.8) would then be applied to transform the item difficulty or the "b" parameter for the test would as follows:

$$= Ab_{(X)1} + B = .8 \times 1.0 + .05 = .85 \qquad (5.11)$$

Once the transformations are completed, the scores can be placed onto a common scale. However, since the linking constants are not known, there are several approaches that can be used to estimate the linking constants needed to place test scores on a common scale. The Stocking-Lord (1983) method used by CTB for the CSAP represents a commonly used approach by test developers for estimating the linking constants. Although full discussion of the Stocking-Lord method is outside the scope of this dissertation, a brief overview of this approach follows to convey the general concept of this method.

Under the Stocking-Lord method, the test characteristic curves (TCC) generated using the common items are compared between two groups of students. A TCC represents the relationship between expected scores on a set of items relative to any given proficiency estimate. The expected score is the sum of the scores that a student would earn across all items given her proficiency level. Continuing with the same example of grade 3 and 4 students, Figure 27 represents an example of TCCs generated for both groups on a set of common items with a sum total of 5 possible points.

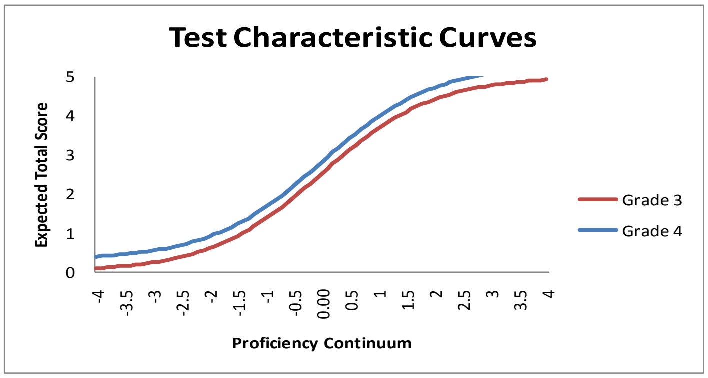


Figure 27. Hypothetical test characteristic curves

In Figure 27, the TCC for grade 4 students is located higher than the TCC for grade 3 students. This finding would not be particularly surprising since, as noted earlier, nothing has been done to adjust the scales to reflect that grade 4 students with expected higher levels of proficiency should exhibit higher scores on the common items than grade 3 students. Under the Stocking-Lord method used by CTB, optimal values or a set of linking constants are chosen to minimize this gap or the distance found between the TCCs. The best set of linking constants that would create overlap between the TCCs is then plugged into Equations (5.7) and (5.8) to transform the item parameters of one test or several tests.

Once the linking constants are estimated and used to transform the item parameters of one test, the scores of both tests can then be placed onto a common scale. In the case of the CSAP, the grade 7 tests in 2001 served as the base grade for transforming the item parameters on every test for every grade[21]. Within the context of the interim assessment, if common items existed between test versions and linking constants were estimated for the version 1 and 3 tests in each grade, the scores from the interim version 3 test could then be transformed to the version 1 test or vice-versa.

This general overview of vertical scaling, particularly in reference to the scaling of the CSAP tests, provides the basis for understanding the procedures conducted in this study to enable performance comparisons across interim test versions 1 and 3. Vertical scaling serves as the only viable method for comparing the performance of students over time on an absolute metric. In the case of the interim assessment, although the groups represent the same set of students, these can be viewed as two separate groups since student proficiency at the beginning of the year should be vastly different from proficiency evaluated at the end of the year. In addition, although both tests are used for the same grade level, these were established as forms with different difficulty levels represented. However, since no common items are present between interim test versions, a set of common items and linking constants were adopted from related assessments as a solution to place the interim test scores on a vertical scale.

The next sub-section describes the underlying assumptions and the process used in this study to enable score comparisons between interim test versions by placing the interim test scores onto the same vertical scale as the CSAP.

Adopting the CSAP Vertical Scales for the Interim Tests

Before placing the interim test scores onto the same scale as the CSAP, the individual item responses from each version 1 test were first combined with the item responses from the prior year's CSAP (in this case, grade 3 data), and the item responses from each version 3 test were combined with responses from the Spring 2008 CSAP item responses (grade 4 data). Placing the interim test scores on the same scale as the CSAP required making two strong assumptions: that the interim test and the CSAP

[21] For each CSAP test, CTB used separate estimation runs referred to as "chain linking" to place more than one test on the same scale as the base grade. The chain linking approach requires estimating the linking constants for each adjacent grade sharing common items and then linking across other grades until the base grade is reached.

assess the same content, and that the administration dates for both interim and corresponding CSAP tests were close enough that the scores generated for either test should be approximately the same for all respondents. In regards to the former assumption, because the CSAP frameworks and power standards are represented on the interim tests, it seems reasonable to assume that both test systems target the same content domain. Further, as discussed at length in previous chapters, the use of these interim tests to predict performance on the CSAP indicates that there is expected correspondence in the content being assessed by both test systems.

Regarding the latter assumption, since there is a time lag between the interim test and corresponding CAP test administration dates, the period of time between the two administration dates may potentially challenge the assumption that the ability levels of students assessed by either CSAP or interim test should be approximately the same. In this study, the assumption is made that a version 1 test would generate similar proficiency estimates to the prior grade CSAP, since, based on existing research, by the time students took the first interim test in the fall, most students may experience a slight loss or decline in grade level proficiency due to the summer break (e.g., Cooper, 1996, Downey et al., 2004). According to one study (Cooper, 1996), summer losses average to approximately 1 to 3 months of grade level equivalent work for most students and this finding further supports the idea that student proficiency by the time the interim test is first taken in the early fall, may be very similar to when student proficiency was assessed back in March.

It is possible that since interim test version 3 is administered approximately a month and a half after the CSAP, the version 3 interim test may be associated with slightly higher proficiency estimates than the CSAP. That is, more learning may have occurred in May when version 3 was administered relative to the time point in March when the same group of students was assessed by the CSAP. Placing the version 3 interim test on the same scale as the earlier administered CSAP, may potentially understate the performance of students on the interim test.

After consolidating the individual item responses for each interim test and corresponding CSAP test, a single estimation run using the same IRT models employed by CTB (the 3-PL and the GPCM) was conducted to generate the item parameters for the interim test and CSAP items. In the single estimation run, the item parameters for the CSAP were held as fixed to those reported in CTB's technical report and the interim test item parameters were allowed to vary. The single estimation run placed the item parameters of the interim test items on the same scale as the selected or accompanying CSAP test. Proficiency estimates (theta levels) on the interim tests were estimated using the unconditional maximum likelihood approach. This final step completed the process of placing the interim test scores onto the same vertical scale of the CSAP. The single run estimation for each interim test version and accompanying CSAP was conducted using the IRT Command Language program (ICL; Hanson 2002) and the linking transformation was computed using the R package plink (Weeks, 2009) in R version 2.9.2 (R Development Core Team, 2009).

Having described the underlying assumptions and process of placing the interim test scores onto the same vertical scale as the CSAP, the next section outlines the approaches used in this study to evaluate two areas pertaining to using the interim tests as teacher effectiveness measures for merit pay. The first area evaluates whether the interim tests can serve as valid measures of growth for teacher effectiveness. This first area compares growth rates achieved by students on the interim tests relative to the CSAP that assess performance over the same content domain. The second area evaluates whether performance outcomes could vary on the mean percent correct score gains and two commonly used SGOs for each classroom when using scores adjusted for the differences in difficulty found across the interim test versions.

Methods

Data

The data set used for these analyses consisted of individual raw item responses from 3,653 grade 4 students who took both versions of the grade 4 reading interim tests, the grade 3 reading CSAP and the grade 4 reading CSAP, and 4,162 grade 4 students who took the math versions of both interim and CSAP tests. Masked student identification numbers were available to match the item response data from the

CSAP with the individual item response data from the interim tests. The district also provided a data set containing a unique teacher identifier for all grade 4 students with accompanying masked student identification numbers. The data sets were matched with the individual item response files using the masked student identification numbers.

Considering that classrooms matched by the teacher identifier varied in size with a minimum of one student and a maximum of 32 students, ProComp staff consulted on September 2009 recommended eliminating classrooms with fewer than ten students from the analyses. The criterion of only including classrooms with at least ten students in the analyses led to the exclusion of 61 classrooms out of a total of 236 classrooms in the district for reading, and 43 out of 235 classrooms for math. According to ProComp staff, they believed that these classrooms with fewer than ten students most likely represented the classrooms of teachers working exclusively with special needs students[22].

Analyses

This section describes three different approaches used to evaluate the inference of whether the interim tests can be used as valid measures of growth to evaluate and reward teachers. The first approach compares the growth made by students in each classroom between the version 1 and version 3 interim tests and between the grades 3 and 4 CSAP tests using a model widely used and referred to in the state as the "Colorado Growth Model" (CGM). The CGM reports the student growth percentiles (SGP) of individual students on the reading, writing, and math CSAP and is used by all districts in the state to annually compare the estimated percentile rank of each student relative to a peer group sharing the same baseline starting point. According to the Colorado Department of Education's SchoolVIEW website, the CGM employs "a statistical model called quantile regression to calculate the student growth percentiles. The calculations use all available test scores to estimate an individual growth score, or student growth percentile. The student growth percentile tells us how a student's current test score compares with that of

[22] For example, these classrooms provide transitional English language services to students or special education pull-out services to students.

other similar students (students across the state whose previous test scores are similar). This process can be understood as a comparison to members of a student's academic peer group." Since the CGM requires a minimum of two waves of data, student growth percentiles (SGPs) for students in grades 4-10 in reading, math and writing CSAP are annually estimated and reported to all districts.

The purpose of employing the CGM in this study is to compare the median[23] growth percentile rank (MGP) associated with a teacher's classroom on each test program. Since the data are restricted to this one district, the academic peer groups in this study are composed of students sharing the same baseline score for each test. The findings from correlating the MGPs between test programs evaluate whether the interim tests are as sensitive as the state tests to assess performance changes taking place in each classroom. A strong correlation found between the MGPs for the interim tests and the CSAP would provide supportive evidence that the interim tests could be used as valid measures of student growth for evaluating and rewarding teachers. The comparison of growth using the CGM entailed evaluating MGPs achieved by classrooms for each test using the original observed raw scores for the interim tests and the CSAP scale scores. The student growth percentiles were calculated using the R SGP package (version 0.0-4) in R version 2.9.2 (R Development Core Team, 2009).

The second approach consisted of a descriptive comparison of the item difficulty parameter for each of the CSAP items relative to the re-calibrated item difficulty parameters of the interim tests. The item difficulty parameter values were plotted to understand the extent to which each interim test version consisted of more or less difficult items than the corresponding CSAP test. This information was used to determine how well the new cuts established in 2007-08 conformed to the level of difficulty found on each interim test.

The third approach consists of two analyses comparing performance outcomes using the *true* and raw scores. The true score is calculated as a function of a respondent's estimated proficiency relative to each item's location. That is, an expected score represents the probability for a given respondent to answer a multiple-choice item correctly (earning a 1) or the higher likelihood of landing in a constructed-

[23] To remain consistent with the state's reporting of CGM generated data, the median SGP is used.

response category. The expected scores are then summed across items to derive a total expected score or "true score" for each respondent. The true scores for the interim tests were generated using the R package plink (Weeks, 2009) in R version 2.9.2 (R Development Core Team, 2009).

The first analysis compares the mean percent correct gains for each classroom using the true and raw scores. For each student, the percent correct gain using each set of scores was calculated by subtracting the percent correct earned on the version 3 test from the percent correct earned on the version 1 test. The percent correct gains were averaged across students to evaluate performance outcomes at the classroom level. The mean percent correct gains for each set of scores were then compared on scatter plots. The purpose of conducting this analysis was to evaluate the extent to which the mean percent correct gains by classroom would differ if using scores adjusted for differences in difficulty found across interim test versions.

The second analysis compared performance outcomes on two commonly used SGO metrics on each test using the true and the raw scores. As mentioned in the beginning of this chapter, two commonly used SGOs in the district are specifying the percentage of students moving up by one performance band, and the percentage of students moving up from below proficient to proficient and above. Since the district's assessment of "growth" based on individual counts of students reaching an objective represents an entirely different metric from the MGPs used in the first set of analyses presented in this chapter, the findings from this analysis are specific to this district but also instructive for other districts who use or are considering the use of SGOs for evaluating teacher performance. That is, the analyses provide additional insight into the extent to which an evaluation of teacher performance can differ depending on what type of scores are used to evaluate the SGOs and depending on how the SGO metric is defined.

Findings

Comparing Growth Achieved using the Colorado Growth Model

Before comparing the median SGPs achieved by each grade 4 classroom for each set of tests, the goodness of fit descriptives (Q-Q Plots) were first examined to evaluate how well the specified theoretical distribution of SGPs fits the observed SGPs. In general, if the quantiles specified for the theoretical and observed data agree, then the plotted points should fall near or on a 45-degree line. Finding generally close alignment between the theoretical and observed data suggests that the model fits the data adequately for estimating percentiles. The Q-Q plots located in Appendix D-2 indicate that in general, the data sets from both test programs fall close to the 45-degree line. Although the plots for the interim test data tend to display more of a "staircase" pattern across the 45 degree line, this pattern is likely due to the fact that in contrast to the large scale score range represented in the CSAP scores, the scores from the interim data have a narrower score range assigned to each test (0 to 32 points possible for reading, 0 to 29 points possible for math). The scatter plots presented in Figure 28 show the strength of the relationship between the MGPs for the reading and math CSAP and interim tests.

In the scatter plots, the 50th percentile point serves as reference lines for evaluating the median SGPs attained, and classroom size is indicated by the color and size of each marker. In Colorado, a median growth percentile (MGP) in the 50th percentile indicates that on average, students made expected or "adequate" learning progress during the academic year. An MGP below the 50th percentile signals lower growth attained by students and an MGP above 50 signals higher levels of growth achieved. The 45 degree line marked on the chart provides another reference point for evaluating differences in MGPs achieved in one test program relative to the other. Points located near or on this 45 degree line indicates classrooms where the MGPs achieved on both test programs were the same or almost equal. The scatter plots in Figure 28 represent MGPs using the original scale of each test (observed scores for the interim tests and scale scores for the CSAP).

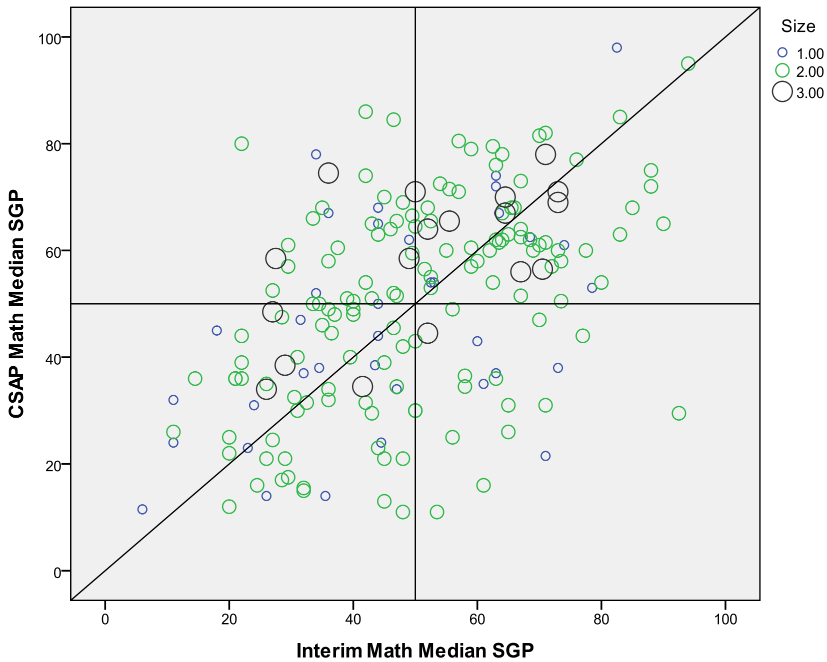


Figure 28. Scatter plots comparing median SGPs by classroom on interim tests and CSAP

In addition to the reference lines, three different class sizes are differentiated on the scatter plots (see legend): a class size between 10 and 15 students were grouped under "1", class sizes of 16 to 25 students were grouped under “2” and classes with more than 25 students were grouped under “3”. These class size groupings serves as another reference point to review the association between classroom size and median SGPs achieved on both tests.

As indicated by the scatters, there appears to be a stronger positive association found for the math tests (r = .48) than the reading tests (r = .33)[24]. However, for both subjects, particularly for reading, these correlations are not strong. Classrooms located in the upper-left hand and lower right-hand regions of the scatter plots represent classrooms where notable performance discrepancies emerged. Correlations of MGPs achieved on each test relative to classroom size are presented in Table 18.

Table 18

Pearson Correlations between Median SGPs for Grade 4 tests
Median SGPs for Math Interim Test and Math CSAP test by Class Group Size

	1	2
Group 1 (*n* =36)		
1. Math Interim Test	—	.51
2. Math CSAP	.51	—
	1	2
Group 2 (*n* =138)		
1. Math Interim Test	—	.48
2. Math CSAP	.48	—
	1	2
Group 3 (*n* =18)		
1. Math Interim Test	—	.6
2. Math CSAP	.6	—

Median SGPs for Reading Interim Test and Reading CSAP test by Class Group Size

	1	2
Group 1 (*n* =50)		
1. Reading Interim Test	—	.15
2. Reading CSAP	.15	—
	1	2
Group 2 (*n* =110)		
1. Reading Interim Test	—	.41
2. Reading CSAP	.41	—
	1	2
Group 3 (*n* = 15)		
1. Reading Interim Test	—	.19
2. Reading CSAP	.15	—

[24] A parallel analysis was conducted that included all classrooms omitted (classes with fewer than 10 students). Including all classrooms with fewer than 10 students resulted in lowering the math correlation from .48 to .39 and increasing the reading correlation from .33 to .35.

As indicated by the data in Table 18, the correlations for math are strongest for the larger classrooms. In reading, however, the correlations are stronger for the group size with the most data or classrooms evaluated (group 2) but are considerably weak in the other two groups.

The histograms presented in Appendix D-3 provide another level of detail showing the magnitude of differences found between the median percentile rank achieved by classrooms for each test. In the histograms, each bar represents a group of classrooms. The bar centered at 0 represents a group of classrooms where small to no differences were found between the MGPs for each test. The bars located to the right of this center bar represent classrooms where the median SGP achieved on the CSAP was larger than the interim test by 10 or more percentile points. The bars located to the left side of the center bar represent classrooms where the MGPs on the interim test was larger than the CSAP by 10 or more percentile points. For Math, there were 76 classrooms or approximately 40% of all grade 4 classrooms in the district where the MGPs for the CSAP exceeded the MGPs for the interim tests by 10 or more percentile points. In addition, a total of 56 or approximately 29% of all classrooms exhibited notably higher SGPs on the interim tests. For reading, the histograms present a similar pattern found on the math tests, but with a considerably higher proportion of classrooms (approximately 50%) exhibiting a median percentile rank on the CSAP that were higher by 10 or more points than the MGPs achieved for the interim tests. Approximately 23% of all classrooms exhibited higher median percentiles on the reading interim tests than on the CSAP.

Since the growth results on the interim tests are tied to monetary compensation, one would hope or expect to find stronger associations between the MGP ranks across classrooms. The extent to which moderate correlations are adequate for making judgments about a teacher's value-added to student learning is a topic of current debate. One recent study released by the Measures of Effective Teaching Project (MET) found similar weak to moderate correlations[25] (r = .377 for math and lower for reading)

[25] This correlation is based on the reported observed correlation in the MET study between the state math test and the balanced math assessment. The disattenuated correlation that is commonly cited from the study is considerably higher at .54.

between supplementary assessments designed to capture higher-order conceptual learning and the state high stakes test and declared that, "teachers who are producing gains on the state tests are generally also promoting deeper conceptual understanding among their students" (Bill & Melinda Gates Foundation, 2010, pg. 9). According to MET researchers, this finding provides strong supportive evidence for calculating a teacher's value-added using the state test since the correlations demonstrate that the state test is associated with student learning of higher-order concepts. However, one critic of the MET study noted that, "while the report's conclusion that teachers who perform well on one measure "tend to" do well on the other is technically correct, the tendency [based on the disattenuated correlations] is shockingly weak…this important result casts substantial doubt on the utility of student test score gains as a measure of teacher effectiveness" (Rothstein, 2011, pg. 3).

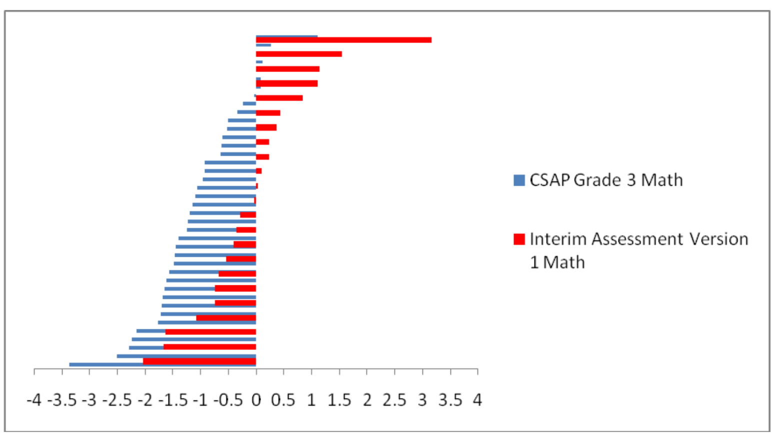
Within the context of this case study, although the DPS interim assessments were designed to be instructionally useful for teachers, the blueprints and the items in the 2007-08 do not indicate that these tests were specifically designed, like the MET study, to capture higher-order learning concepts. That is, the lower correlations found on the MET study may be deemed acceptable or high enough since those assessments possess more depth in content relative to the breadth of content represented on the state test. However, for the interim tests, since the CSAP framework statements are embedded in each test, higher correlations between the MGPs achieved should be expected.

The next set of analysis compares the item difficulty of the interim tests relative to the CSAP. The information from this analysis provides insight into the extent to which the interim tests are reflective of more easy or difficult items than each corresponding CSAP test.

Comparing Item Difficulty across Tests

This sub-section presents findings from comparing the item difficulty parameter values of each interim test version relative to the accompanying CSAP. These comparisons were conducted to learn if there were any clear differences in the level of difficulty found between each interim test and corresponding CSAP test and whether these differences have any implications for the cut-scores set by DPS staff to define student proficiency. The comparisons of item difficulty are depicted in the bar charts located in Figure 29 and Figure 30. In each bar chart, each bar represents an item and the length of the bar represents the difficulty of that item. The items in each chart are sorted by difficulty (from easy to difficult). Negative values indicate easier items and positive values indicate items that are more difficult. For example, for the first chart in Figure 31, the items range in difficulty from approximately just below -3.5 to close to 3.5 logits. The chart indicates that the grade 3 CSAP math test is populated with easier items than the version 1 interim math test since there are many more items with higher negative values that exceed the negative values on the interim math test. In addition, the chart indicates that the interim math test had one item that was extremely difficult since that item was located close to 4 logits.

The two bar charts presented in Figure 29 below show the range of item difficulties for all CR and MC items for the math tests. Item difficulties across response categories for CR items were averaged to yield a single item difficulty value.



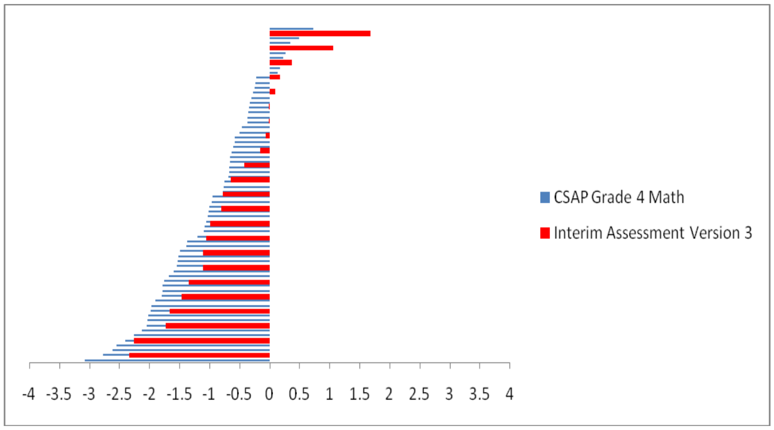
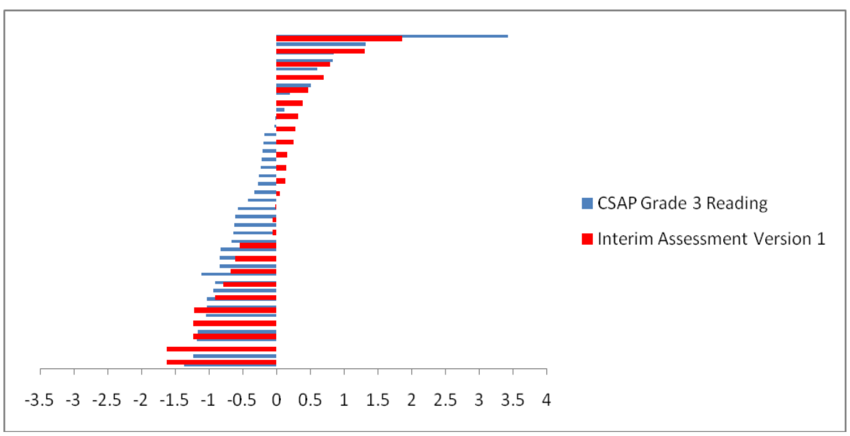


Figure 29. Comparisons of Item Difficulty on Math Tests

The bar charts presented in Figure 29 indicate that both interim math tests were populated by more difficult items than the comparison CSAP tests. The average difficulty for the interim math version 1 items relative to the grade 3 CSAP items is -.04 compared to -1.03 on the CSAP. The average difficulty for the interim math version 3 items relative to the CSAP math items is -.64 compared to -.99 on the CSAP. The bar charts in Figure 30 compare the item difficulty on the reading tests on the interim assessments and the CSAP.



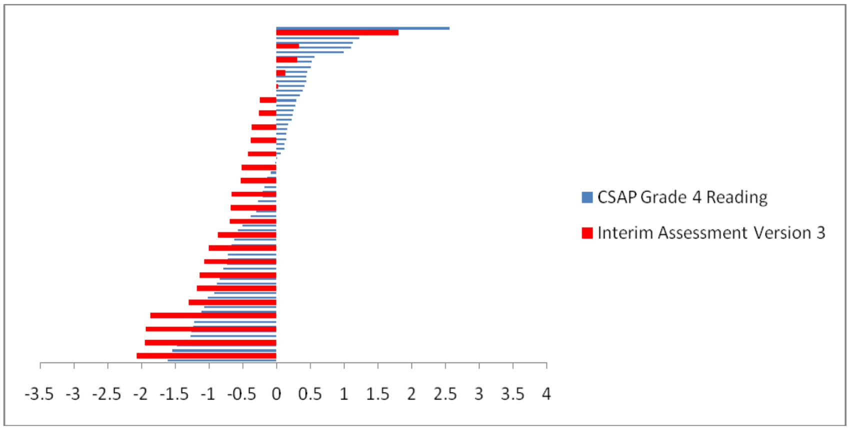


Figure 30. Comparisons of Item Difficulty on Reading Tests

The bar charts above indicate that although the version 1 interim reading test was more difficult on average than the corresponding grade 3 CSAP reading test, the grade 4 CSAP reading test consisted of more difficult items than the version 3 interim reading test. The average difficulty found on the grade 3 CSAP reading test was -.29 compared to -.08 on the version 1 interim reading test. The average difficulty found on the grade 4 CSAP reading was -.24 compared to -.66 on the version 3 interim reading test.

To evaluate how well the cut-scores set conformed to the difficulty level found on each interim test, the proficiency classifications of all students who took the grade 4 CSAP reading and math tests and the interim version 3 reading and math tests were compared. Table 19 presents a comparison of how students are classified on the grade 4 CSAP relative to each interim version 3 test reviewed.

Table 19

Comparison of performance classifications for math

	Grade 4 CSAP Math				
Grade 4 version 3 Math Test	A	P	PP	U	Total
Advanced	525	165	4	0	694
Proficient	308	1061	339	15	1723
Partially Proficient	17	366	947	254	1584
Unsatisfactory	1	20	273	732	1026

Comparison of performance classifications for reading

	Grade 4 CSAP Reading				
Grade 4 Version 3 Reading Test	A	P	PP	U	Total
Advanced	101	482	17	0	600
Proficient	29	1444	1041	122	2636
Partially Proficient	0	62	536	541	1139
Unsatisfactory	0	17	74	547	638

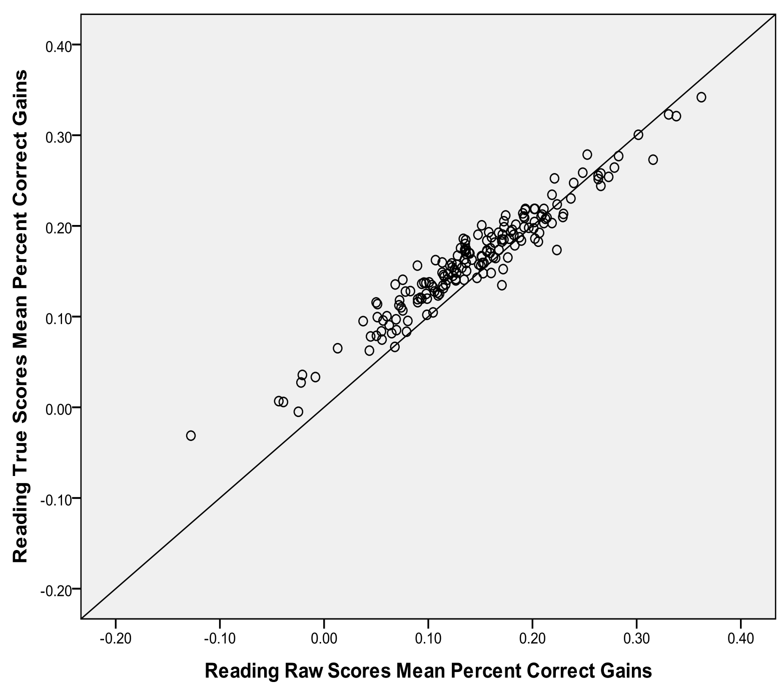
In the tables, all values located on the diagonal indicate agreement in the performance classification of students on both test programs. All highlighted values located in the off-diagonals indicate where students earned different classifications across tests. As indicated by the table for reading, the new set of cuts resulted in classifying approximately 44% of all students at a higher proficiency level on the interim test compared to the CSAP. For math, the new cuts for the interim tests appeared to be set too high for the advanced region. As indicated by Table 19, approximately 44% of students who scored "advanced" on the grade 4 Math CSAP were classified as "proficient" students. The findings from this sub-section reveal that despite best efforts to set cuts that more accurately reflected the level of difficulty found on each test, these cuts should be reviewed if the objective is to better align the proficiency classifications of students on these tests with the CSAP. However, as pointed out in Chapter 4, district staff should also consider whether there are a sufficient number of items on the interim tests to support using four proficiency classifications.

The last set of analyses evaluates performance outcomes at the classroom level using two sets of scores: the true scores and the raw scores. The purpose of conducting these analyses was to evaluate whether performance outcomes in classrooms may vary when comparing the mean percent correct gains or the percentage of students meeting each SGO, if using scores adjusted for differences in difficulty found across interim test versions.

Comparing Classroom Growth Outcomes Using the Raw and Adjusted Scores

Comparing Raw and True Score Mean Percent Correct Gains

In this analysis, the mean percent correct gains for each classroom using the true scores and the raw scores were compared to evaluate the extent to which performance outcomes may differ when adjusting the scores to account for differences in difficulty found across interim test versions. The scatter plots presented in Figure 31 show the differences in the mean percent correct gains found for the reading and the math tests using both sets of scores.



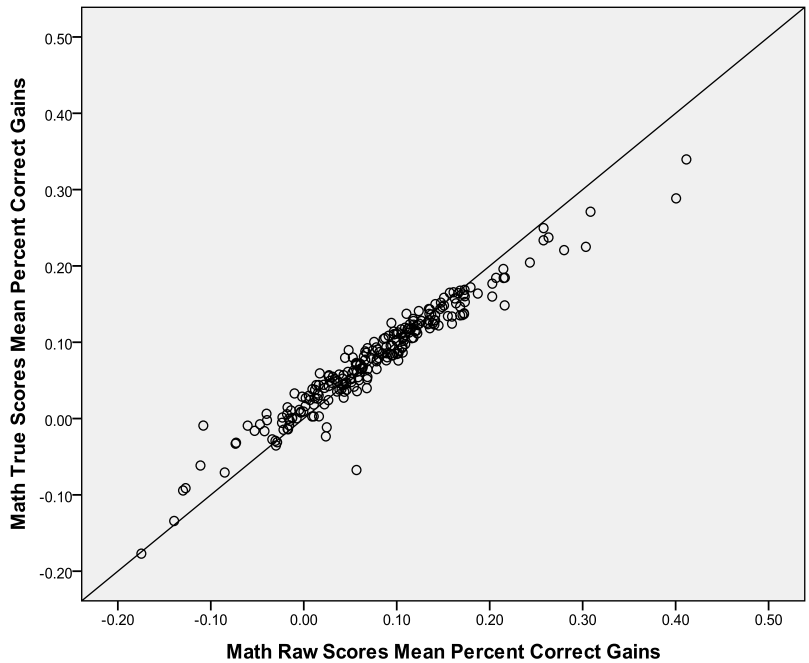


Figure 31. Scatter plots of mean percent correct gains by classroom for reading and math interim tests

In Figure 31, the points located on, or near, the 45-degree line in each scatter plot represent those classrooms where the mean percent correct gains were similar or identical (if on the line) based on either set of scores. For the reading tests, the first scatter plot indicates that the majority of classrooms have higher mean percent correct gains when using the true scores. Table 20 presents the mean, standard deviation, minimum and maximum for the reading tests using both sets of scores. As indicated in Table 20, the true scores on average, yield higher mean percent correct gain scores than the raw scores. In addition, 79% of classrooms had higher mean percent correct gain scores for reading using the true scores.

Table 20

Comparison of Mean Percent Correct Gain Scores by Classroom for Reading

	Mean Percent Correct Gain	
	Raw Scores	True Scores
Mean	.14	.16
Standard Deviation	.07	.06
Minimum	-.13	-.03
Maximum	.36	.34
% of Classrooms with Higher Gains	.21	.79

For the math tests, the second scatter plot indicates that the gains were similar for many classrooms using either set of scores. As indicated by Table 21, the mean, standard deviation, and minimum for the math tests using both sets of scores reveal similar results.

Table 21

Comparison of Mean Percent Correct Gain Scores by Classroom for Math

	Mean Percent Correct Gain	
	Raw Scores	True Scores
Mean	.083	.085
Standard Deviation	.08	.07
Minimum	-.14	-.13
Maximum	.41	.34
% of Classrooms with Higher Gains	.42	.58

However, when comparing the number of classrooms that earned a higher gain score using either set of scores, 58% of classrooms showed higher mean percent correct gain scores on math when using the true scores.

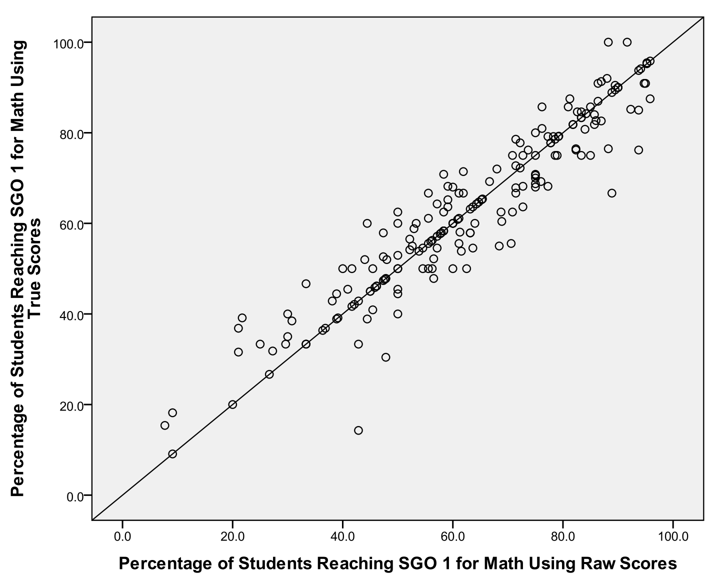
Overall, the findings reveal that depending on the set of scores used to assess student performance, the mean percent correct gains are largely higher when using the true scores. This finding suggests that the raw scores are likely understating the performance achieved by students in grade 4 classrooms, especially in reading, and shows the importance of adjusting the scores to account for differences in difficulty found across interim tests versions.

Comparing Performance Band Movements

In this last analysis, the percentage of students by grade 4 classroom reaching each commonly set SGOs was computed using the true and the raw scores. As indicated earlier, two commonly set SGOs were: 1) setting a target based on the percentage of students moving up by one performance band and 2) setting a target based on the percentage of students reaching proficiency by interim test version 3.

This analysis compared the differences in the percentages of students reaching each type of SGO by classroom using each set of scores. The percentages of students reaching each objective using both true and raw scores are presented on the scatter plots in Figures 36 and 36.

The scatter plots in Figures 32 and 33 represent the percentage of students by classroom who met each SGO. The first scatter plot in each figure presents the percentage of students in each classroom who moved up by at least one performance band by the end of the school year. The second scatter plot presents the percentage of students in each classroom who moved from below proficient to proficient by the end of the school year. In each scatter, the classrooms located on the 45-degree line represent classrooms where no differences in percentages were found using either set of scores. Classrooms located below the line represent classrooms where teachers were favored by the observed scores. For the classrooms favored by the observed scores, this finding indicates some students in these classrooms were assigned a lower true score based on how they responded to items with varying characteristics (e.g., easier or more difficult items, or items that discriminated more or less between students). As noted earlier, differences found between the raw and true score performance can differ since the true score reflects how a student performed across different types of items.



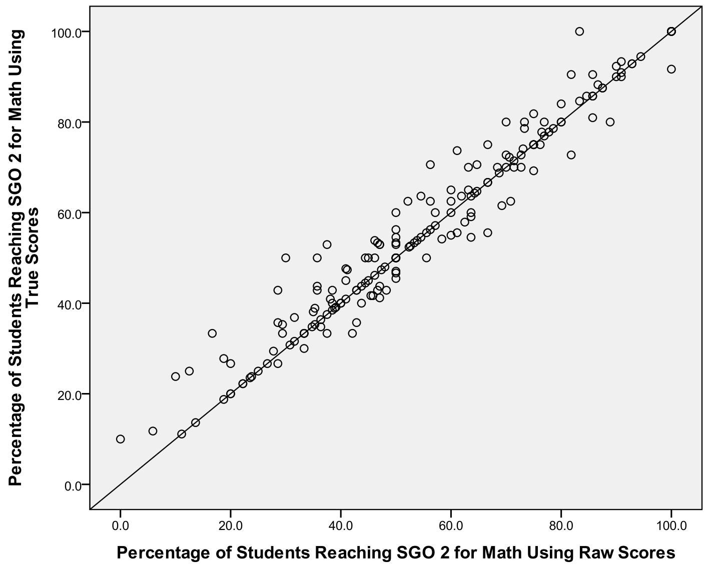
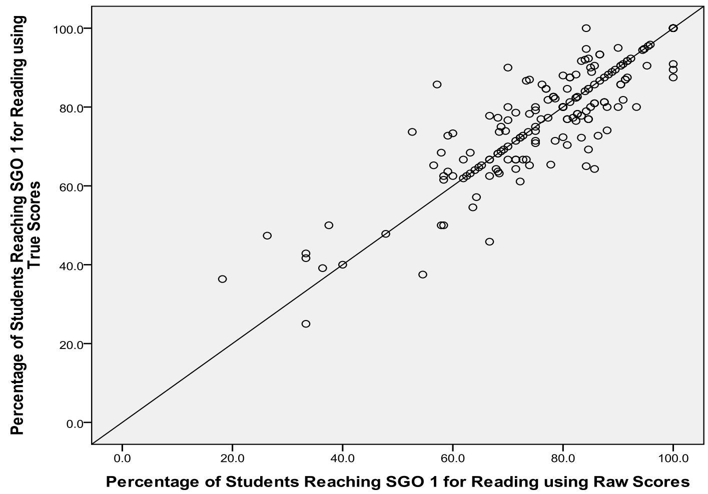


Figure 32. Comparisons of the percentage of students reaching SGOs 1 and 2 based on true and raw scores for math tests

The scatter plots for math in Figure 32 indicate that there were many classrooms exhibiting similar SGO outcomes based on either set of scores. These classrooms are situated either on or very close to the 45-degree line. However, some differences emerged for a considerable number of classrooms. The first scatter plot in Figure 32 suggests that a similar number of classrooms had higher percentages of students making the first growth objective on either set of scores. That is, for some classrooms, the performance of some students was favored using the raw scores. For other classrooms, the performance of students was favored using the true scores. The second scatter plot in Figure 32 suggests that there appeared to be more classrooms exhibiting higher percentages of students meeting the new SGO using the true scores relative to the raw scores. That is, as indicated by the figure, there were more classrooms located above the 45-degree line than there were classrooms located below the line. In addition, the classrooms located above the 45-degree line are more scattered or spread out relative to the classrooms located below the line. Since this objective is tied to the percentage of students moving up from below proficient to above proficient, this finding would suggest that more classrooms would exhibit higher percentages of students reaching proficient based on these true scores. The scatter plots for the reading tests are presented in Figure 33.



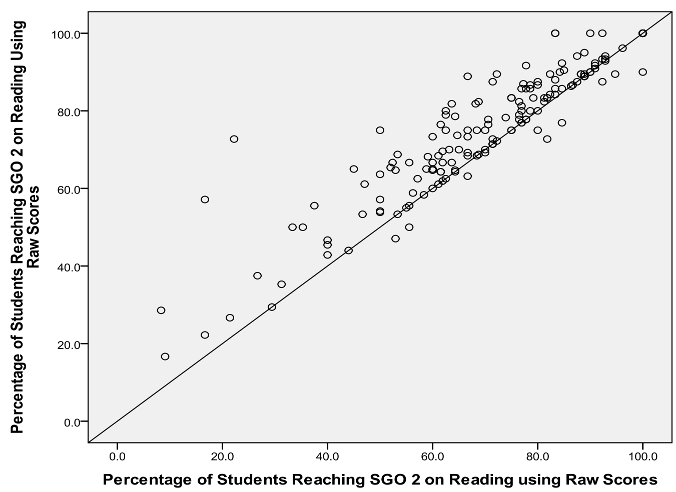


Figure 33. Comparisons of the percentage of students reaching SGOs 1 and 2 based on true and raw scores for reading tests

The scatter plots for the reading tests present findings that in comparison to math, many more classrooms would exhibit substantially higher gains on the true scores on the second SGO, but that the gains would vary for each classroom on the first SGO depending on which scores are used. Similar to math, the first scatter plot in Figure 33 suggests that a similar number of classrooms exhibited higher gains on one set of scores relative to the other. However, in contrast to the math test, the second scatter plot in Figure 33 suggests that substantially more classrooms yield higher percentages of students reaching the second SGO objective on the reading test when using the true scores.

The findings from comparing the raw and true score performance of each classroom in Figures 32 and 33 suggest that different outcomes would have likely been achieved by several classrooms if the scores were adjusted to account for difficulty found across test versions. As indicated by the scatter plots, although some classrooms exhibited higher percentages based on the raw scores, many more classrooms exhibited higher percentages on the true scores when the objective was crafted towards evaluating the percentage of students who moved from below proficient to proficient.

Summary

Taken together, the findings from this chapter indicates that the assumption that these interim tests can serve as valid growth measures for evaluating teacher performance and merit pay is problematic. Despite the considerable improvements made by staff to improve the tests for the 2007-08 school year. The comparison of median growth percentiles achieved on the interim tests relative to the CSAP highlighted discrepancies in the growth percentile ranks achieved in each classroom. These discrepancies were more pronounced for the reading tests where more classrooms appeared to achieve higher MGPs on the CSAP than on the interim reading tests. As noted earlier, considering that both CSAP and interim tests are designed to evaluate student performance on state standards, and the CSAP framework statements are embedded in the blueprints of the interim tests, one would anticipate finding higher associations on the growth percentiles on each test rather than the moderate correlations found. The findings from comparing the median SGPs achieved between test programs have larger implications for the current and growing trend across districts and states encouraging the use of interim assessments as measures of teacher effectiveness in both state tested and non-state tested subjects. These implications are discussed in the final chapter.

The analyses comparing the mean percent correct gains and SGO outcomes using the true and raw scores on the interim tests reveal that the underlying properties of the scores used and the definition of the SGO metric could result in different performance outcomes for many teachers. The findings from evaluating outcomes using two different sets of scores and two commonly used SGOs in this district, have implications for states and districts who use or are planning to use SGOs for evaluating teacher effectiveness[26]. As seen in these analyses, the true scores yielded higher mean percent correct gains for the majority of all classrooms on the reading tests and for 56% of all classrooms on the math tests. Further, the design of the SGO metric can also largely influence how teachers are evaluated and rewarded. These findings imply that districts who use or are planning to use SGOs in teacher evaluations need to consider not only the design of the SGO, but should also consider the properties of the underlying scores being used to evaluate student growth.

A final point for consideration is that the weak to moderate correlations found between the estimated achievement of students on tests measuring students over the same content domain in this study raises questions of whether one set of assessments (interim assessments) can be used to evaluate student growth for teacher performance. A similar concern was raised by Lockwood, McCaffrey, Hamilton, Stecher, Le and Martinez (2007) who found that notably different estimates of teacher effectiveness were found across various subscores and the weighting of those subscores on a single math assessment. For example, if a teacher emphasizes her teaching in a sub-content area that is not highly weighted by the test, then that teacher's effectiveness rating or value-added score would be significantly lower than if she had emphasized a highly weighted sub-content area. Since different effectiveness ratings could be attributed to teachers on a single test, Lockwood et al. caution against interpreting teacher value-added estimates from assessments as pure, stable measures of teacher effectiveness. That is, a teacher's value-added estimate (e.g., such as the median SGP earned) could be influenced by factors such as how much weight on a test is placed on a given standard or subscore. For this reason, Lockwood et al. recommend that

[26] Louisiana serves as an example of a state planning to have all teachers instructing non-state tested subjects develop SGOs as measures of teacher effectiveness.

assessments used for value-added purposes should be closely aligned to the scope and sequence used in the classrooms to ensure the scores reflected are a good match for the teacher (pg. 21). The assumptions supporting the first segment of the interpretive argument not evaluated in this dissertation (content design) would need to be checked to establish whether the interim assessments adequately represent content taught by the teacher during the school year.

The next chapter presents the final set of findings in this study, which evaluates an entirely different set of inferences and assumptions. The evaluation of the claim that the interim assessments can be used by teachers to improve their instruction was based largely on findings from interviewing a small group of veteran teachers in this school district.

CHAPTER 6 - Evaluating Teacher Use of Interim Assessments to Improve Instruction

The last set of analyses presented in this study was conducted to evaluate the claim made by DPS staff that teachers can use interim assessments to improve instruction to meet the needs of individual students. This claim about how interim assessments can improve instructional practices and subsequently enhance student learning and outcomes represents a commonly articulated claim from the testing industry. For example, CTB claims on their website that, "Acuity® puts a robust set of interim and formative solutions into your classrooms—and the right support can help make it easier for your teachers to spot learning gaps and address them quickly with targeted instruction." Another testing company, the Northwest Evaluation Association states on their website that, "NWEA classroom reports offer a host of features to help teachers put the data to use quickly...both growth and proficiency information is included to quickly diagnose student needs and make instructional decisions where they have the greatest impact." The assumptions and inferences used to evaluate the claim of whether teachers can use the interim assessments to improve their instructional practices was shown earlier in Figure 8.

Figure 8 begins with two assumptions about the interim tests. Since the studies in Chapters 4 and 5 checked the assumptions under the first inference in Figure 8, this chapter focuses on evaluating the assumptions under Inferences 2 through 4. In this chapter, the evaluation of the interpretive argument proceeds despite the finding that the evidence from the previous chapters appears to challenge the assumptions under the first inference. That is, the findings from the two previous chapters suggested that a student's proficiency on the interim test may not correspond to the proficiency classification of the same student on the CSAP and that some sub-scores on the interim assessments did not have adequate reliability for assessing an individual student's performance (e.g., $\alpha < .5$). The implications of this finding means that although stakeholders may have assumed that the interim tests provided accurate classifications of student performance and that diagnostic information about individual students could be obtained from the subscores, the findings from the previous chapter suggest that these assumptions are not supported by strong evidence.

The evaluation of the interpretive argument in this chapter draws on a set of interview and survey data to evaluate the extent to which teachers use the interim assessments to improve instruction. The study first examined whether teachers were able to make key connections with the interim assessment data to inform their instructional practices. The key connections, as noted under Assumption 2.1, are that teachers can evaluate a student's responses to the test and use that information to identify gaps in student learning and conceptual misunderstanding with content taught prior to each interim test administration date. Evidence supporting the third assumption (3.1.) provides information on the extent to which teachers can incorporate the information from the interim assessments into their existing contextual framework for each student and the extent to which these data illuminate where teachers need to provide targeted instruction for students. Evidence addressing the fourth assumption was gathered to evaluate the extent to which teachers can then use the contextual framework enhanced or potentially improved by the interim assessment data to address conceptual errors and provide better instructional supports to students.

Two set of exploratory analyses are presented in this chapter. The first analysis examines teacher responses to a district-wide survey to understand the extent to which they value interim assessments for gaining useful information about students and the frequency in which they review or use these data. The second analysis draws on evidence from the perspectives of DPS veteran teachers to examine the extent to which those teachers believe the interim tests can locate student learning gaps and misconception with content taught in class, and the extent to which they can act on the information to improve their instruction. The next section describes the methods for evaluating the claim that the interim assessments are useful for teachers to improve instruction.

Methods

Procedure for Administering Survey

Survey items about the benchmark and course assessments were included on a district-wide electronic survey that was administered by electronic mail to all DPS teachers at the end of the 2007-08 school year. The survey was developed by external researchers studying the ProComp system and was sent through the Office of the Chief Academic Officer in DPS to all teachers in May of 2008. Teachers were allowed to submit their responses during the summer break. The total population of teachers sampled for the survey was 4,537. This figure, provided by DPS ProComp staff, represents the total number of teachers employed in the district during the month of May 2008. The total number of respondents to the survey was 2,207 or 49% of the total population of teachers in DPS.

Sampling Procedure Used for Selecting Interviewed Teachers

A purposive sample of eight master teachers was selected from a pool of twenty candidates recommended by the DPS central staff to participate in one-hour long interviews. DPS staff identified these teachers as "master teachers" or teachers recognized for their deep content knowledge and commitment to the teaching profession. More importantly, DPS staff selected these teachers since each of these teachers was recognized as having cultivated the practice or routine of using data to drive their instructional practices.

The sample of teachers selected for this study taught at different levels (elementary, middle, and high) and taught in schools with different state accountability performance designations. Table 22 presents information on the school levels and accountability designations represented by teachers in the sample.

Table 22

State Accountability Designation of School (in 2007-08) and Level by Teacher

Teacher	Accountability Designation	Level
1	Low	Elementary
2	Low	Elementary
3	Excellent	Elementary
4	Low	Middle
5	Average	Middle
6	Low	High
7	Low	High
8	Average	High

To ensure that data from different types of schools were included, schools with different accountability designations were selected. The sample was restricted to master teachers to ensure that information about the usefulness of the interim assessment data would not be influenced by feedback from less experienced and new teachers unaccustomed to both teaching and to the use of data to inform their instruction. Years of experience in the teaching profession ranged in this group from a minimum of nine years to 40 years. This wide range of teaching experience found for the group of interviewed teachers was not an intentional part of this study's design. All teachers interviewed taught in the district before, and during the time that the interim assessments were first implemented in the 2006-07 school year. The teachers provided feedback based on their past two years and current year of using the tests in the classroom.

Survey Development and Evaluating Survey Responses

To evaluate the perspectives and uses of the interim assessment district-wide, three items developed and used by a researcher studying the use of interim assessments in a school district (Neild, 2008) were included in a district-wide survey to all DPS teachers. The items included in the district-wide survey are shown in Appendix E-2. In Neild's (2008) study, these items were developed to understand the extent to which schools and teachers were dedicating time and practice around using the interim assessment data. More specifically, Neild wanted to evaluate whether the instructional communities in each school were using the interim assessments during professional development and whether teachers believed that they were gaining instructionally useful insights about their students and their own teaching practices from these assessments. Responses to the three items shown in Appendix E-2 provided a district-wide perspective into how often teachers in DPS reviewed the information from the tests and whether teachers believed they gained new knowledge either about students or about their instruction from these tests. Descriptive statistics were used in this study to summarize the frequency of responses for each survey item.

Procedures for Conducting Teacher Interviews and Evaluating Interview Data

All interviews with teachers consisted of open-ended dialogues to determine whether teachers with considerable teaching experience used the assessment to improve their instruction. As noted earlier in Chapter 1, test developers and school districts typically cite evidence from research on formative assessments (e.g., Black and William, 1998) to substantiate the use of interim assessments to improve instructional practices and student achievement. In Black and William's seminal literature review of formative assessments, teachers used the assessments to diagnose student misconceptions with content, to gain feedback on the effectiveness of instructional strategies deployed, to improve instruction, and to develop strategies with students to establish learning objectives or targets. Although there are many aspects of formative assessment practices that can be evaluated, the interpretive argument presented earlier in Figure 8 is limited to the instructional improvement piece.

To evaluate the extent to which instructional practices may have been improved by data from the interim assessments, veteran teachers were asked to describe the type of activities and interactions that take place in the classroom using the interim assessments. The following prompts initiated the conversation with each teacher during the one-hour long interview session: What do the interim assessments tell you about your students? How are you using this information in the classroom (for 2008-09 data)? In what ways have you used the 2007-08 data?

The decision to use an open-ended interview format was motivated by advice received from a researcher participating in a CRESST study evaluating how teachers use interim assessments in the classrooms. According to the researcher, she observed that teachers appeared to be tailoring responses based on what they believed the formal interview protocol was trying to elicit. According to the researcher, "we felt that teachers were telling us what we wanted to hear; so it was really hard to get a feel for whether they were truly using the [interim] assessments in the classroom or just simply telling us that they were using these assessments because the protocol gave them information in advance that that is what we were trying to learn" (Frohbieter, personal communication, 2008). Although an open-ended interview format in-itself may not prevent subjects from conforming their responses to what they believe the researcher is seeking to learn, teachers were asked to give specific examples of how they were using the data and to provide examples of how that data related back to their instructional practice. For example, if a teacher indicated that they used interim assessment data to "adjust instruction", the teacher was asked to provide examples of how they interpreted the data and what aspects of the data were used to adjust their classroom practices.

All interviews were recorded and transcribed within five days of the interview to a Word document. The Word documents were then exported to NVIVO 8 and analyzed using two traditional coding schemes. The first coding scheme comprised of codes developed apriori and was based entirely on examining whether data supporting Inferences 2-4 emerged across all interviews. Data points providing information on how teachers could use the assessment to identify areas where students needed intervention or more assistance and to help students know what they need to do in order to improve were highlighted to support assumption 2.1. Data points providing supportive evidence that teachers could use the assessment to learn about whether teaching strategies were effective in improving student mastery over content taught in the classroom were highlighted to support assumption 3.1. Data points illuminating whether teachers were improving their instructional delivery based on information gleaned from the assessment were highlighted to support assumption 4.1. The code for each of these three areas created before undertaking the first review of the interview transcripts is noted in the first column in Appendix E-1. After sections of transcript data were sorted using the pre-determined set of codes, a second review of the transcripts was conducted under an inductive approach designed to explore whether additional themes or patterns emerged across the interviews. The second column in Appendix E-1 also shows the new codes that were generated during the second review of the transcripts. After the transcript data were coded under the two coding schemes, the data were then analyzed in NVIVO to document the frequencies of patterns and responses emerging from the interview data.

The codes highlighted in Appendix E-1 represent areas where common and recurring themes emerged across interviews and gave direct insight into the extent to which each of the assumptions under the interpretive argument for this study holds. Three of the highlighted codes in green in Appendix E-1 were treated as relational codes under a larger theme called "Understanding the relationship between instruction and content standards". This larger theme provided the most data illuminating the extent to which all three of the assumptions hold. In addition to the theme exploring the relationship between instruction and the content standards, three additional themes were explored to evaluate whether the actions taken with the interim assessment data appear to conform to the expected use of these assessments to improve instruction.

Findings

Based on the procedures described in the previous section, the exploratory findings presented in this section provide evidence on the extent to which the interim assessments appear to be used by teachers for improving instruction to meet individual student needs. This section begins by highlighting key findings from the summary of responses to the district-wide survey, then presents findings from interviewing DPS veteran teachers. The survey results were evaluated to determine the extent to which the data supports Inferences 2 and 3 in Figure 1. In this study, the assumption was made that if the second and the third inferences did not hold, the likelihood that the interim assessments were being used to improve instructional practices was low. For the teacher interviews, the findings are organized according to larger themes that emerged from the interview data. Three of the four themes explored provide supportive evidence on the extent to which each of the Inferences hold. The last theme reflected a larger perspective shared across teachers on the current number of assessments used in the district to monitor student learning.

Findings from Survey Data

The survey yielded a response rate of approximately 49% of the total teacher population and was distributed to every DPS school teacher who taught during the 2007-08 school year. The survey results provided in Appendix E-2 shows the frequencies around each response category for each item. As seen in the percentages reported for each response category, the findings from the survey were mixed. As indicated by the survey results, on average, approximately 45% of all teachers indicated that they found the assessments to be useful for evaluating their students and roughly, the same proportion of teachers indicated that their schools are using the data for planning purposes. On average, approximately 26% of all teachers in the population did not form an opinion (selected the neutral category) about the assessments that were designed to help them and approximately 29% of teachers "disagreed" with the utility of the interim assessments for the classroom.

Although there were no open ended questions asked of DPS teachers, 117 or approximately 5% of teachers surveyed chose to comment on the interim assessments when asked on the survey whether they had "any other comments to share". All teachers who elected to provide feedback generally framed their opinions and input about the interim assessment system with questions about the system's utility. The following quotes reflect a sample of feedback received from this group of teachers:

"Benchmarks are simply a reiteration of the information teachers gather in the process of examining student work."

"Benchmarks are graded by teachers who want their students to do well - is this a reliable measure?"

"What is the cost/benefit analysis of the benchmark assessment? Are benchmark assessments cost effective?"

We were already making assessments and grouping. The Benchmarks take a huge amount of time and are not that much more beneficial than what seasoned teachers already do in their practice.

"We test our kids too much. Benchmark, CSAP prep, SRI Reading take too much focus away from the classroom and its demands."

"I am a SPED teacher. I evaluate my students' skills on a regular basis. I look at the benchmark test results, but have other more detailed assessments."

Although this feedback comes from a small group of teachers and may not be representative of the views of the larger population of teachers, the information is presented to not exclude any voices or concerns being articulated by one segment of the population. In addition, this feedback would resonate with the teachers interviewed for this project since they largely used other assessments and artifacts to identify student misconceptions.

The findings from the survey data suggest that less than half of all teachers surveyed responded positively overall to the items on the survey. Although district envisioned that these assessments would help teachers improve their instructional practices, the survey results indicate that approximately half of the population of teachers surveyed was most likely not integrating the interim assessments data for instructional improvement purposes.

Findings from Teacher Interviews

To understand and document how teachers use interim assessment data, all teachers were asked to provide examples of what they did with the data when the scoring reports became available and how or whether they used the data for any purpose they deemed useful. Although each teacher offered unique perspectives about the interim assessments, four common themes emerged across the interviews and captured how teachers used the data, whether the teachers could identify student misconceptions and strengths using the data, and the extent to which interim assessment data could be used to improve their instruction for individual students.

Theme 1: Understanding the relationship between instruction and content standards.

A key point made by all interviewed teachers is that the interim assessments serve as a "useful tool" for gaining general knowledge about the content standards and how students perform on specific standards. However, since the interviewed teachers emphasized that they were already familiar with state standards, they indicated that they believed less experienced teachers benefited from these tests more than experienced teachers. According to these teachers, less experienced teachers would stand to benefit more from the interim assessments, since these tests explicitly outline or make apparent, standards that are emphasized both in the curriculum and in the CSAP. The following data points across interviews speak to the utility of the interim assessments for evaluating student performance on particular standards:

"…what I try to explain to [parents] is, I try to say 'Your student is currently unsatisfactory and this is the beginning of the year test and by the end of the year, by benchmark number three, we want them to be totally proficient on all of the sixth grade standards.' And then we go through and we look at the information by each standard."

"…this student, he's a seventh grader. He's partially proficient on this standard. So, I would ask, 'what's he doing here and what do we need to work on more with writing?' So, yes, sometimes I do take a look at the standards and I might write a goal to go along with, you know, if it was a reading goal, maybe reading standard one, because he's lower there. I…look at my other tests and say 'Okay, well he's low here, and what are some other examples of that?'

"We've identified [from the benchmarks] that we need to work on main idea. We're starting to focus on a specific standard now where we're doing two weeks teaching a specific skill and then we're going to assess it again, then we're going to start into some month long cycles and do different ones."

"So this tells us in this group, only 9% of these kids were proficient. 50% of them were unsatisfactory. Here is the breakdown by different standards. So we can look at that and we can discover that on this one, standard four, actually 50% of the class did get over 50% of the points possible. So that's pretty telling. So maybe I don't need to spend quite so much time on standard four."

"We don't know why, but we have looked, and some of my teachers have done more of that than I have, looked at the district data and concluded we need to focus more on short, constructed response activities to have our students do better for this year. We have a huge focus there this year. So that's directly related to the test and the benchmark…hopefully we looked at the data correctly and we're hoping we'll see some results."

The above data points suggest that teachers use the interim assessment data to evaluate student performance over each standard. In some cases, several teachers indicated that they also looked at how students performed on each CSAP framework statement represented on the test. The data points also suggest that at some school sites, instructional decisions appear to be influenced by evaluating how students perform on the standards measured by each test. Despite their stated familiarity with state standards, the data points suggest that these veteran teachers appear to use the interim assessments as a check to identify whether their students either as a group or on an individual basis, display mastery or weakness on the standards represented on the test.

Although all eight of the teachers framed the tool as "useful", particularly due to its focus on the standards represented on the high-stakes CSAP tests and its alignment with standards represented in the curriculum, the teachers did not describe or provide examples of using the interim assessments as a tool for diagnosing student misconceptions. In other words, the tool appeared to provide information on what areas (standards) they believed they needed to focus on in class, but was not described by teachers as a tool that can illuminate student misconceptions with content, or help students establish learning targets. For example, although teachers indicated that the tool, "*…showed me that all students had trouble with understanding poetry…but we hadn't spent much time on poetry*" or "*…I learned that I forgot to teach [the standard] that addresses research skills since not one of my students got that one (item) correct*", and that "*…I can see learning gaps when I see that my students missed certain standards that I probably didn't focus on in the classroom*", these comments and similar data points gathered from teachers point to how the tool is useful to them for identifying areas where instruction on a specific standard was predominantly absent. When each teacher was asked specifically to give examples of how the instrument helped them identify student misconceptions, comments regarding the utility of the tool when "*all students*" or "*most students*" missed an item underlying a framework statement or did poorly on a specific standard re-surfaced. No examples could be found in the interview transcripts where teachers used the data to identify conceptual errors, but rather, this group of veteran teachers consistently emphasized the use of these assessments in their classroom practice for illuminating the performance of their students on specific state standards.

Because this group of teachers did not use these assessments to identify student conceptual errors, their ability to use these assessments to provide targeted instruction to meet individual student needs appeared limited to modifying instruction to re-teach and address standards in the classroom. When the teachers were asked to give examples of how they used the interim data to provide targeted instruction to students, rather than to simply re-teach standards not previously or adequately covered in class, the responses received from this group of teachers moved away from discussing standards. As indicated by several teachers, the interim assessment "*isn't designed to give teachers insights into how we need to modify our instruction*" or "*to directly answer your question, I think it's somewhat correct to say that probably there isn't a high amount of insight from it or high level information gained from it for instruction*" or "*but in terms of you know, is it giving me any information that I could do to improve my daily instruction for my students? No*". In other words, the tool provided these teachers with information that convinced them that they needed to augment their lesson plans to include specific standards that they may not typically focus on, but did not appear to be providing them with information that illuminates how to meet the instructional needs of individual students. During the interviews, all teachers also noted that in general, unless most or all of their students appeared to miss a specific standard on an interim test, they would prefer to not act on the interim assessment data for individual students.

Theme 2: Triangulating across assessments to evaluate student learning

To understand whether interim tests were used by these teachers in meet the individual learning needs of student, teachers were asked to describe what sources of information they relied on to help them understand how best to meet individual student need. In each interview, teachers made references to different activities and sources, but no reference was made by these teachers of incorporating the interim assessment data to help improve of facilitate their understanding of student learning. All of the teachers established that they triangulated across multiple artifacts and assessments before deciding whether an individual student is truly struggling with specific content. The following data speak to how these teachers draw on information about students from different sources:

"My every day teaching itself is diagnostic…so a lot of it I rely on my own ability to do it and use so many different pieces of evidence… I think if you're an exceptional teacher and you're doing the diagnostics every day yourself and you know you're moving students along and you're doing everything to make sure that happens, it would nullify. That's kind of what's implied in your question. It might nullify the constant checking. What is going on with these kids?"

"*I don't have to go back to a database some place always to know what this student demonstrates and to know what I need to do to help them get better. I don't have to spend a ton of time on it because it's a given. The way I teach, I know. I find out every day when I look at papers and I see missing parts and can see that a student couldn't understand this…we worked with a detail chart today, which I use to see what's filled in and what's not. In other words, what they're getting and what they're not.*"

"I use a lot of different assessments to look at children and where they're reading and writing….And I interview the kids and I do one on one conferencing with them. I also hand out a little survey asking – just trying to find out what student attitudes are toward reading and writing. So you know I use a lot of different assessment pieces --in order to kind of figure out where each of my students are."

"Because some of the things they [the items] ask – that you are looking at in either of those [interim reading] tests really don't give me valuable information. I get much more valuable information from the DRA (Direct Reading Assessment)."

"I mean as we're reading along, reading is such a hand on the pulse thing that you're reading along with your students, you can see where they're at and it's like if the wound is bleeding, you're not going to go back and look at [the benchmarks]. You're going to take care of that wound right there and that's what reading is for me."

The selected data points above provide examples of how this group of teachers depended largely on other assessments or activities with students to diagnose student misconceptions. As indicated earlier, these teachers appeared to value the interim assessments for providing them with insights on how their students performed on state standards, but did not reference the use of the interim assessment data as a data point for modifying their instruction to meet individual student needs. For this group of teachers, the interview data suggest that instructional actions associated with the interim test data were motivated largely by classroom performance on standards. Based on the data points explored under the first theme and how teachers triangulate across other sources to locate student misconceptions, it would appear that the uses of the interim assessment for this group of veteran teachers did not support Inferences 2-4 specified in the interpretive argument (see Figure 8). The interview data suggest that teachers evaluated the interim assessment data to review how their classes performed on standards, but utilized other sources to understand how best to tailor their instruction to meet individual learning needs. These findings, however, do not mean that other teachers in the district do not use the interim tests to identify student misconceptions with content. In fact, five of the eight interviewed teachers voiced concern that because the district is populated with novice and new teachers[27] and is impacted by teacher turnover, many inexperienced teachers may over-value the data from interim assessments and spend an inordinate amount of time trying to re-teach information that is only supported by a "*few items.*" This concern is also shared by Shepard (2007) who noted that interim tests that are often designed to address multiple areas within a domain could result in, "a long list of discrete skill deficiencies requiring inexperienced teachers to give 1,000 mini-lessons".

---

[27] In the specific case of one interviewed teacher, only two other teachers in the entire school had more than two years of teaching experience.

Theme 3: Gaining insight into testing behaviors.

In addition to valuing the interim tests for providing them with data on how their students perform on standards, the teachers indicated that the instrument provides substantial insights into testing behavior prior to the high-stakes administration of the CSAP in the Spring. This theme emerged across the interviews frequently and the data points below speak to the testing behavior insights gained from administering the interim assessments:

"I get to see who would probably sleep during the CSAP, who gets nervous, who gets sick to their stomach, who decides to not show up because it's a formal testing day….I even get surprises like students who I know know the [content] really well but who get physically ill about these tests!"

"It gives me a better sense of who [should get] an accommodation and that's good to know because it's better to know earlier on before the CSAP."

"The good thing about these [interim assessments] is that in the past, we didn't know how students would act around taking the CSAP and this really gives us an opportunity to figure out who will have problems taking the CSAP because we treat it just like the CSAP…it's much much shorter, but we go through the same serious and formal process of administering these tests to students so that [the students] take it seriously and do their best."

"Even though we're a high performing school, we still have students who aren't test takers and get really nervous about testing and when we test them [on the benchmarks] we get to see what we can do for them before the CSAP. It's hard though, because I have little ones (third graders) who don't like having to take these tests being separated from their friends and they become really intimidated by the whole process…"

"*Well, I think going back really quick to the idea of practicing a test, this is very different from anything we do in the classrooms because what we do in these days is we actually do it as if it were a CSAP day. So, we actually get into groups where – you know, let me give you an example. Some of my students might have a reading accommodation or they might have an extended time, and when we're doing inclusion in the regular class, when we're in here, I can do that on a normal basis with the kids in that group. But, then when CSAP time comes around, everybody's grouped according to their accommodations, so I have a whole group of people that we read the test to at test times to, versus a whole group of extended time. And I think it's good to get them into that mindset that they're in a different environment and to be around other people that aren't always in their class…"*

The data points captured above suggest that teachers appeared to appreciate being able to better predict and potentially manage student reactions and accommodations during the testing window in preparation for the high-stakes CSAP. The extent to which teachers appreciated knowing student testing behaviors before the high-stakes testing period was surprising, particularly since this theme emerged even for those teachers who were teaching in higher performing schools.

Because this finding was unexpected but emerged as a dominant theme across interviews, informal feedback was requested from assessment expert, Dr. Lorrie Shepard, to better understand why teachers would place value into gaining insight into testing behavior. Dr. Shepard indicated that this finding was not particularly surprising within the context of high-stakes testing, and that the teachers were most likely interested in gaining any kind of information that would help them better predict student outcomes (which is also influenced by testing behaviors) on these assessments (personal communication, 2008).

One study using a narrative approach to better comprehend the views of educators on testing related to accountability (Craig, 2004) supports this notion raised by Dr. Shepard regarding how the psyche and practices of educators are greatly shaped and influenced by accountability pressures. In this study, Craig (2004) collected narrative data over multiple years to learn how the educational landscape and views of educators changed in an urban high school facing great pressure to increase student achievement results. At the high school site, Craig found that:

> To [the principal's] way of thinking, a sustained focus on curriculum and instruction will lead to academic growth on the part of students, enhanced professionalism on the part of teachers, and incremental change with respect to standardized test scores. These three phenomena, however, struggle to coexist with the ''daily drip'' method because that approach demands that teachers repeatedly administer the same antidote─ more and better versions of test preparation─ to students. When this occurs, neither teachers' nor students' prior knowledge is taken into account. And over time, testing-taking addictions set in as opposed to a robust relationship between and among curriculum, instruction, and assessment. And scores related to testing, rather than the human learning they represent, take front and center stage. (pg. 8)

This notion described by Craig of how scores and testing take "front and center stage" despite best efforts by some teachers and principals to make a concerted effort to focus on classroom activities potentially illuminates why most teachers in this sample may have felt that gaining insight into testing behavior was valuable to them. That is, since the productivity of these teachers are being defined largely by a constellation of accountability systems, any specific piece of information that would help these teachers better predict student outcomes on these tests would be deemed valuable.

Theme 4: Diminishing returns from multiple formal assessments.

The second theme reviewed earlier revealed that all eight of the teachers triangulated across different assessments and information sources to evaluate their students. Although these teachers were highly supportive of the use of multiple assessments, seven of the eight teachers interviewed mentioned that there were many other formal assessments in place and that the number of assessments being used to evaluate performance needs to be prioritized and potentially narrowed down to a more coherent system. It is important to note here that not one of the teachers interviewed argued against testing their students during the school year but rather, expressed disappointment with the investment in multiple assessments that were not necessarily adding any new information to what they already understood about their students. That is, many of these other assessments did not appear to enrich the contextual framework teachers had about each student.

The following data points speak to this theme of teachers wanting a more coherent structure in the system for assessing their students:

"…the whole testing thing is, as I say, I'm speaking as an extreme supporter of testing and as an extreme critic of the way it's become. It's become very cumbersome."

"We're getting an overload of data, some is useful, some not quite as useful – but it's – I think what is happening is that there's so much that there's almost too much..."[28]

"No, we don't need two weeks [of test preparation] for CSAP. We really don't need the time we're spending right now with these two or three days or a week for these district things. Maybe working every day with snippets of information throughout a painless method for teachers and kids it would get better, more efficient, smarter information."

For the interviewed teachers managing information from different interim assessment systems, the discordant pieces of information appeared to result in diminishing and in some cases, negative returns for gaining information about their students. Two out of seven teachers who expressed disappointment with the current body of assessments administered in the district noted that they would "not mind"

[28] In the specific case of the teacher captured in this quote, multiple literacy assessments were being implemented to fulfill the requirements of a grant funder; and these assessments were producing data, in her opinion, that no teachers in the building believed accurately assessed their students.

diverting resources from the interim assessments to professional development, but both also stated that the interim assessments provided them with better information than other administered assessments. Three other teachers identified school specific interim assessments that they believed did not add value to their understanding of student performance. For these teachers, the assessments identified as not providing useful information to them were mandated by donors supporting specific initiatives at those schools.

In summary, the findings from these interviews suggest that the interim assessments provided useful information to these teachers on how students were performing on content standards measured by the CSAP tests and student testing behaviors. However, the data from the interviews do not suggest that this group of teachers rely on these instruments to identify how they need to improve their instruction to meet individual student needs or to address common student misconceptions. As indicated by the findings under Themes 1 and 2, teachers consistently described the interim assessments as very useful at the standards level. More specifically, the teachers indicated that they appreciated reviewing how their students perform on state standards at each time point, that the tests made the standards more apparent across the district, and in the case of some teachers, that the tests prompted discussions with colleagues on which standards require more instructional focus across classrooms. Although the data connected to the standards appear to be valued by these teachers, this information does provide teachers with conceptual insights into student learning.

Out of the eight teachers interviewed, only one teacher provided a specific example of using the interim tests as a tool for interacting with students. The other seven teachers stated that they did not go over the assessment results with students but would review the results with parents or with the department head or principal. When this teacher was asked to elaborate on these data conversations with students, she pointed to a chart marking where individual student fell in each proficiency band for each standard and said, *"after we get the [interim] test results, I update the chart and have everyone look at it together so that we can see and discuss as a class how many more students need to move up to proficient on each of the standard being measured…the students who aren't proficient know who they are and they know* *that they're letting their classmates down if they don't reach our proficient goals*". This type of feedback given to students by the teacher does not align with the dynamic richness of the continuous feedback loops taking between teacher and individual students described by the formative assessment literature. That is, within the context of this example, the feedback does not appear to be focused on the content and concepts that need to be covered to reach proficiency, but rather, illuminate and focus on just pointing out the distance between a student's current achievement and the proficient bar.

The findings from these interviews do not suggest that dynamic and rich feedback loops are not taking place within classrooms or that these teachers are not engaging in formative practices, but rather, suggest that the interim assessments are not being utilized by these teachers a tool for improving instruction. The next section review the findings from a survey used to collect district-wide feedback from teachers about the interim assessment system.

Summary

Although the interviews in this study were conducted with a very small sample of teachers, the findings from these interviews largely parallel the findings in the larger set of interviews conducted by Olah et al. (2010)[29]. Teachers in this study and in Olah et al.'s study valued the data provided by the interim assessments, but their responses do not indicate that teachers can gain conceptual insights about student learning from these assessments. In this study, the interviewed teachers used the assessments to modify their instruction to re-teach standards to the class. Further, teachers presented examples of identifying student conceptual errors by drawing on other sources of evidence and did not reference the interim assessments as a key source for enriching their understanding of student learning. Based on the interviews with this group of veteran teachers, it would appear that Inferences 2 through 4 specified in the interpretive argument presented earlier in Figure 8 would not be supported. However, the consideration remains open that these assumptions may hold if a different group or type (e.g., novice teachers) of teachers were interviewed.

[29] These studies were reviewed earlier in Chapter 3.

The survey response data evaluated here present findings similar to those found in the survey results from Clune and White's (2008) study. Both Clune and White's study and the survey findings present mixed evidence of teachers integrating the interim assessment data in their regular classroom practices. That is, both studies found that teacher use of interim data appear to fall between two groups: those who used the data independently and in discussion with others, and those who seldom or never reviewed the interim assessment data.

Overall, the interview data suggest that the assessments used during the time that the items came directly from TPR lack content depth to provide diagnostic information about students. However, even if diagnostic data to improve instruction may have been gleaned from these tests, findings from both the interview and survey data suggest that information on how to use these data as a conduit for illuminating teaching practices and student knowledge about the content domain being tested needs to be made more apparent. The findings from the district-wide survey in particular, point to the possible absence of opportunities for learning how to use the interim assessments during the first two years of implementation. Although the evidence gathered in this study was restricted to evaluating how well the assumptions noted in Figure 8 held, DPS assessment staff were also asked to provide feedback on whether professional development opportunities or training was taking place across the district to provide guidance on how to interpret and use the data for instructional improvement.

According to DPS assessment staff, most of the activities over the past three years of implementing these tests went into developing the tests and ensuring the proper administration of the test and very little time was dedicated to training teachers on using data from these tests (personal communication, 2009, 2010). In other words, assessment staff did not have the opportunity to work with assessment leaders at each school site to ensure that teachers knew how to use data from the assessments and to develop strategies that would help them strengthen student learning around content measured by the test. During the 2009-2010 year, district staff intended to roll out a focused professional development program to help teachers interpret and use these data for the classrooms. A follow-up study would need to be conducted to evaluate whether profession development has resulted in the desirable outcome of seeing teachers actively use the interim data to inform and adjust their instructional practices. In addition, a follow-up study should be conducted to evaluate whether the current set of interim tests consists of items that reflect more depth in content than the set of interim assessments evaluated in this dissertation.

CHAPTER 7 - Discussion

Despite the large investment and rapid deployment of interim assessments in school districts across the nation, the variability of standards used to develop these tests, and the expectation by users that these assessments provide valid data for evaluative, predictive, and instructional uses, few studies have been conducted to examine whether specific uses of the test can be supported or justified (Perie et al, 2009; Herman et al., 2008; Shepard, 2007; Herman & Baker, 2005). Shepard (2009) states, "interim assessments could sometimes be a good thing, but they are brand new and wholly unexamined. Therefore, some amount of skepticism and search for evidence is warranted" (pg. 35). Although measurement experts beginning with Cronbach (1988) and later with Shepard (2007, 1993), Kane (2006, 1992) and Wilson (2004) have long advocated for the use of an evaluative framework to validate tests, limited analysis has been conducted at the district or state level to establish policy or procedures to evaluate and validate the uses of interim tests and the claims made about what these tests can achieve.

The primary purpose of this dissertation is to evaluate whether the uses of an interim assessment program are warranted. Because the experience of co-creating interim assessments with a test vendor and the uses evaluated in this study are not unique to this one school district, the findings from this study likely apply to the experiences of other school districts across the country. Two of the four uses evaluated in this dissertation represent commonly documented uses of interim assessments at a national level as tools to predict outcomes on high-stakes summative state tests and to improve classroom instruction. The use of the interim assessments for merit pay purposes is gaining more popularity nation-wide as more districts and state departments of education consider using interim assessments (particularly for non-state tested subjects) to evaluate teacher performance for merit pay. The fourth use evaluated in this study, the one-time use of interim assessments to identify students for mandatory summer remediation in DPS, has been documented in other school districts such as in New York City. Although using interim assessments to identify students for mandatory summer remediation may not represent a common use at the national level, a more common use of interim assessments that could impose similar high consequences for students is the use of interim assessments to form ability groups or to assign grades (Niemi, Vallone, Wang and Griffin, 2007).

The analyses in this dissertation evaluated two sets of interpretive arguments represented in the figures below.

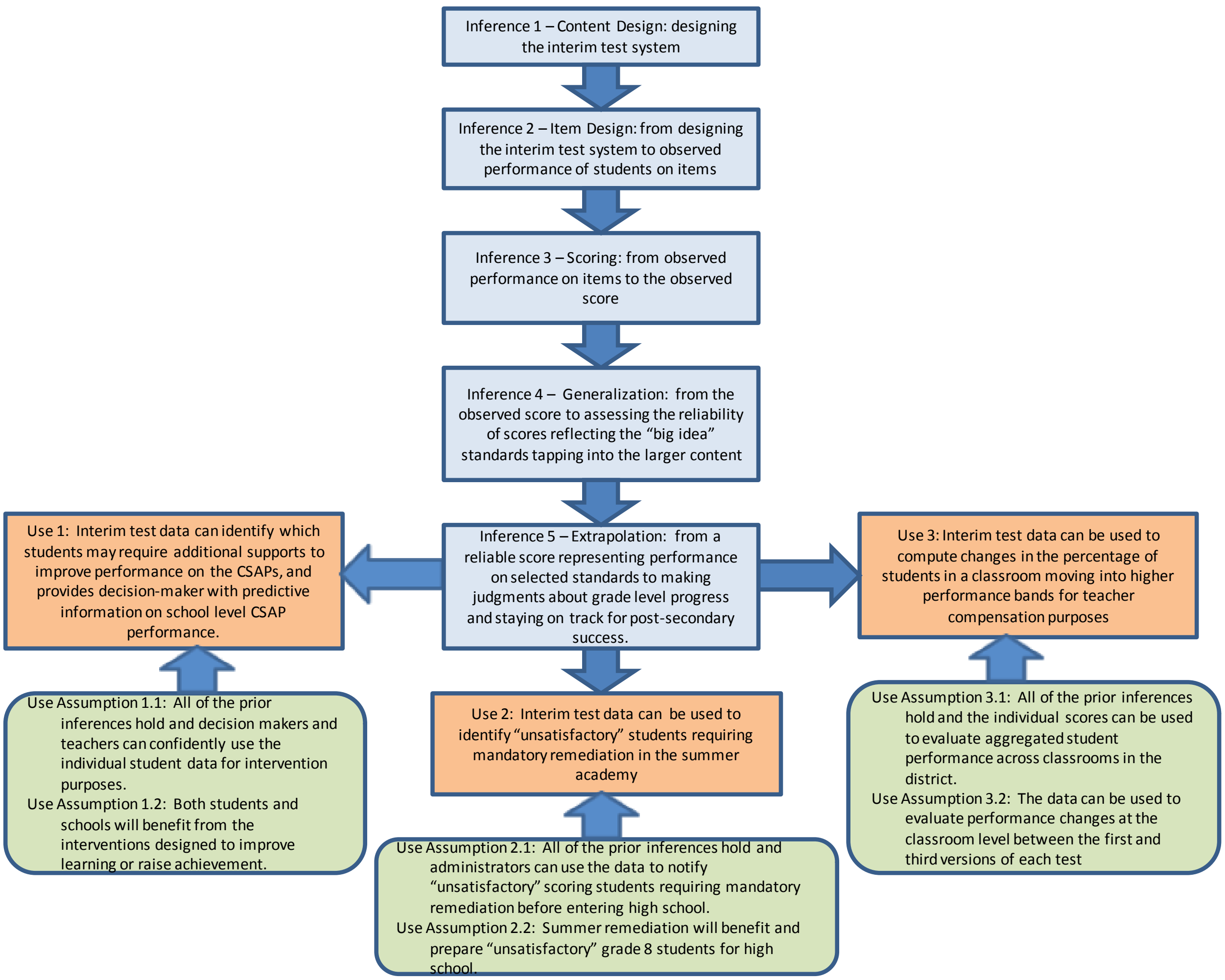


Figure 34. Interpretive argument for using interim assessments to support predictive, mandatory remediation and merit pay purposes

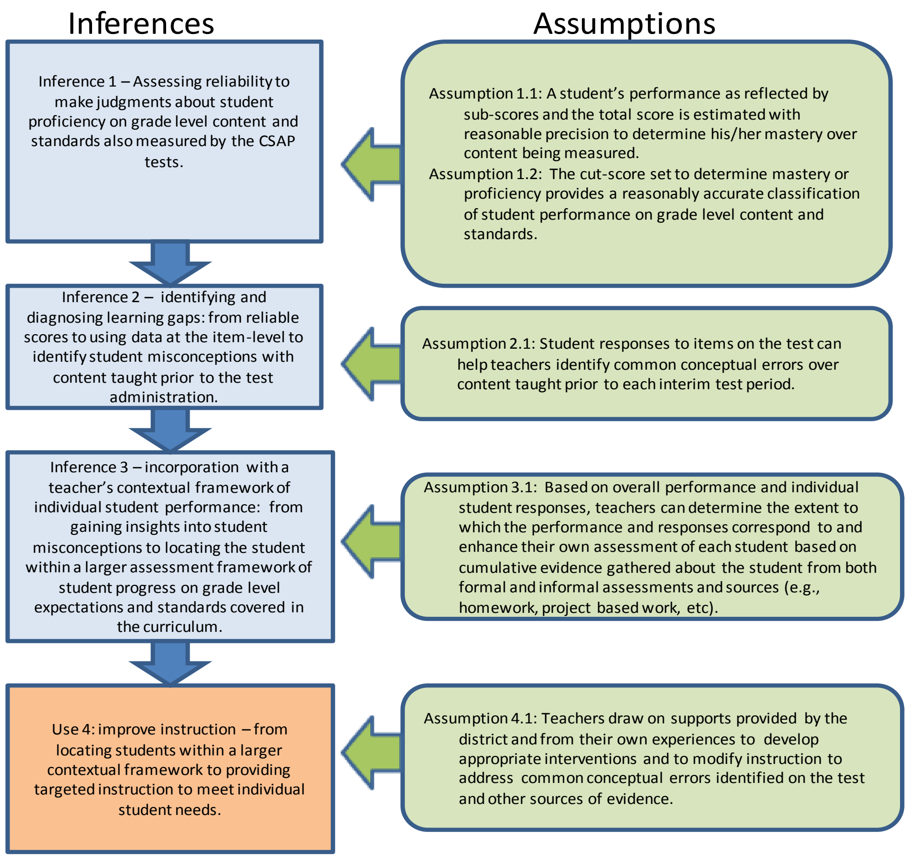


Figure 35. Interpretive argument for using interim assessments to improve instruction

In this chapter, I discuss the extent to which each of the four uses presented in these figures is supported by adequate evidence from evaluating the supportive inferences and assumptions, along with implications of the dissertation findings for other school districts and the broader educational field. Following the discussion of the findings and implications associated with the evaluated uses, a section on future research based partly on the acknowledged limitations in this dissertation identifies additional areas of research that would contribute to the existing set of studies evaluating interim assessments and the associated uses.

The Validity Argument

Predictive Uses

Overall, for all nine of the assessments evaluated using classical test statistics, the findings revealed that the tests exhibited reasonably adequate levels of reliability, largely consisted of items with acceptable p-value ranges and point-biserials, and appeared strongly correlated to the CSAP tests. The strong correlations found of above .7 for eight out of nine tests are impressive when considering the limited number of items represented on each test. Although the analyses for the Scoring Inference revealed that there were one to two multiple-choice items on each test that exhibited point-biserials lower than .3 and at least one constructed response item on each test with reversed point-biserials between a higher and a lower response category, these items did not appear to have a large effect on each test's reliability. Based on these qualities, these tests appear to provide the predictive information needed to identify which schools, classrooms or students will exhibit higher or lower levels of performance on the CSAP. Considering that most district staff value this predictive use highly, the interim tests appear to provide adequate information about expected performance. Further, considering the compressed timeframe given to item panel members to develop these interim tests and the small number of items available to assess student mastery over content tied to the "power" standards represented on the CSAP test, it is remarkable that district staff were able to accomplish this objective of creating predictive assessments with limited time and resources. Based on the findings from Chapter 4, it appears that this predictive use can be justified.

Mandatory Remediation

Although the interim tests are strongly correlated with the CSAP and classical test statistics largely point to items that meet technical standards, elevating the stakes of the test for mandatory remediation during the summer of 2007 required evaluating the cut-points set for these tests more closely. As indicated in Chapter 2, in 2006-07, a best-guess estimate was used to set the cut-points for all interim tests administered that year. In 2007-08, a new method based on matching or modifying items to conform to set scoring ranges was used to improve upon the cuts set during the previous year. For the grade 8 math test used for mandatory remediation, and for other interim assessments evaluated in both years, the conditional standard error of measurement (SEM) generated under an IRT approach for the cut point where most items were located was approximately .4 or higher. When considering the small range of points attributed to each of the four proficiency levels and a SEM of .4, the likelihood of misclassifying students in different performance bands increases. In the case of the grade 8 math test, approximately a third of all students could fall into one or two performance bands when applying a 67% confidence interval. Further, when compared to the CSAP, 38% of the students classified as unsatisfactory on the interim tests were located in other proficiency levels on the CSAP during the same year.

Despite the fact that the process used in 2007-08 to match or modify items to each proficiency level represented a vast improvement over the process used in 2006-07 to set cut-points, a considerable number of students were likely misclassified on the 2007-08 assessments and the current set of interim assessments. These misclassifications would likely occur since panel members may have misjudged whether an item could be answered correctly by a specific type (e.g., proficient or partially proficient) of student and a small or limited range of points are still defined for each of the four proficiency levels. As seen in the comparison of performance classifications on the grade 4 interim tests with the CSAP in Chapter 5, a large number of students were either classified higher or lower on the reading and math interim tests relative to the CSAP. Due to the considerable uncertainty found in locating students in a specific performance band, it appears that the use of the proficiency level information on the interim tests to drive decisions that have higher consequences for students, such as mandatory remediation, would not be supported by these tests.

Evaluating and Rewarding Teachers

In contrast to the first two uses, the use of this test to drive merit pay decisions entailed comparing performance gains made by students between two interim test versions. Although the student growth objectives (SGOs) set by teachers using these assessments typically compared performance band movements across tests, the primary focus of the analyses in Chapter 5 was to evaluate whether assumptions about using these tests to evaluate growth are supported. Evaluating whether the interim tests can serve as valid measures of student growth entailed comparing the growth achieved by students on these tests relative to the high-stakes CSAP tests. Using a set of grade 4 tests administered in the 2007-08 year, the findings revealed that comparing growth as expressed by the median growth percentiles (MGPs) and using the original raw scores earned by students, yielded weak (for reading) to moderate (for math) levels of correlations.

Since the gains made by students on these interim tests are tied to monetary compensation, the correlations found in comparing growth rates achieved by the same group of students across test programs in this study is disconcerting. Since the MGP ranking of classrooms varied for the same group of students on the interim tests relative to the CSAP, these findings raise the question at a broader national level of whether interim assessments can be used to fairly evaluate teacher effectiveness in classrooms. That is, these findings suggest that district and state entities selecting an interim assessment program to evaluate teacher performance should first evaluate how well the chosen test assesses student growth prior to assuming that these tests provide valid measures for assessing student growth. The question of fairness is particularly relevant today since many districts and states propose evaluating student growth using interim assessments for teachers instructing non-state tested subjects.

In addition to the growth comparisons, in the case of this district and other school districts that use interim assessments to evaluate teacher performance using student growth objectives (SGOs), the findings from Chapter 5 illuminate the importance of considering the properties of the underlying scores used and the design of the SGO metric. As indicated by the findings from comparing the mean percent correct gains and each SGO assessed using the true and the raw scores, the performance outcomes varied for classrooms depending on the set of scores reviewed and the type of SGO used. Overall, the findings in Chapter 5 question the validity of these assessments as measures of growth and of teacher effectiveness.

Instructional Improvement

The fourth use evaluated in this dissertation was evaluated largely through qualitative evidence on whether interim assessments could improve instruction from the perspective of veteran teachers. The survey data indicated that on average, approximately 45% of surveyed teachers responded positively to questions on the survey. The interviewed veteran teachers placed value on the interim assessments, but not one of the eight teachers interviewed described their use of the assessments as a tool for improving instruction to meet the needs of individual students. As indicated by the findings from Chapter 6, teachers used the interim assessments to identify if they failed to teach, or adequately address, a standard for all students in the classroom. However, the interviewed teachers indicated that the tests did not provide them with information to assess individual student strengths and weaknesses over content taught previously and the extent to which they needed to adjust their instruction to accommodate the needs of individual students.

To date, the district claims on their website that the interim assessments “provide formative information for teachers to tailor their instruction based on the individual needs of their students”. However, based on the feedback from the small group of veteran teachers and the survey data, it appears that the assessments play a role in the classrooms that are distinct from documented formative practices (see Black and William, 1998). Despite the fact that the findings from this study came largely from the perspective of a small group of veteran teachers, these findings largely parallel findings from other studies conducted with larger groups of teachers and reviewed in Chapter 3. Although it is possible that other groups of teachers, such as novice teachers, may be using these tests to improve instruction for students, the extent to which the data from the tests are providing conceptual insights about student learning remains questionable.

Validity Argument Summary and Implications

To summarize the validity argument, each figure presented below list the strengths and criticisms associated with each use evaluated in this study. As indicated in each of the four figures below, the predictive use of these interim tests appears supported by the identified strengths, but all other uses appear largely unsupported by the evidence gathered in this study.

Use 1	Strengths	Criticisms
Predictive or Early Warning System	**Inference 2** • Interims consist of items with varying difficulty to evaluate students at various levels of proficiency **Inference 3** • Most items display acceptable p-values and point-biserials **Inference 4** • Tests strongly correlated with the CSAP (.7 and above for 8 out of 9 tests) • Most tests are reliable at .8 and above	*If decisions for predictive uses depend on performance band information:* **Inference 2** • Insufficient number of items to support the use of four performance bands **Inference 4** • Due to insufficient number of items to support the use of four performance bands, many students could potentially be misclassified into 2 or 3 performance bands (depending on which confidence interval is applied) • Performance band cuts for the interim assessments have either been set too high or too low (depending on content area and grade) relative to the CSAP

Figure 36. Evaluating predictive use: summary of strengths and weaknesses.

Based on the strengths and weaknesses of each inference summarized above, predictive uses fulfilling low-stakes purposes are justified. However because the criticisms noted under inferences 2 and 4 invalidate predictive uses between the performance bands set by the interim tests relative to the CSAP performance bands, inferences about student performance band outcomes on the CSAP based on data from these interim tests would also be largely unwarranted.

<table>
<tr><th>Use 2</th><th>Strengths</th><th>Criticisms</th></tr>
<tr><td>Mandatory summer remediation (one-time use) or any decision that has high consequences for individual students using the interim tests (e.g., leveling students at the beginning of the school year)</td><td>Inference 2
Interims consist of items with varying difficulty to evaluate students at various levels of proficiency
Inference 3
Most items display acceptable p-values and point-biserials
Inference 4
Tests strongly correlated with the CSAP (.7 and above for 8 out of 9 tests)
Most tests are reliable at .8 and above</td><td>Inference 2
Insufficient number of items to support the use of four performance bands
Inference 4
Due to insufficient number of items to support the use of four performance bands, many students could potentially be misclassified into 2 or 3 performance bands (depending on which confidence interval is applied)
Performance band cuts for the interim assessments have either been set too high or too low (depending on content area and grade) relative to the CSAP</td></tr>
</table>

Figure 37. Evaluating mandatory remediation: summary of strengths and weaknesses.

Based largely on the weak evidence found supporting Inference 4, the application of the interim test results to inform decisions with higher consequences for individual students is not justified by this test. Further, even if this inference were to be strengthened (e.g., using only one cut point to differentiate lower performing from higher performing students), using one test to drive decisions that have high consequences for students cannot be justified under any circumstance. A test can only evaluate students over a sub-set of a given content domain and is therefore inherently limited in its capacity to precisely and accurately evaluate a student’s set of knowledge and skills. Decisions that impose higher consequences on students should always be triangulated using different sources of information.

Use 3	Strengths	Criticisms
Teacher merit pay	**Inference 2** • Interims consist of items with varying difficulty to evaluate students at various levels of proficiency **Inference 3** • Most items display acceptable p-values and point-biserials **Inference 4** • Tests strongly correlated with the CSAP (.7 and above for 8 out of 9 tests) • Most tests are reliable at .8 and above	**Inference 2** • Insufficient number of items to support the use of four performance bands **Inference 4** • Due to insufficient number of items to support the use of four performance bands, many students could potentially be misclassified into 2 or 3 performance bands (depending on which confidence interval is applied) • Performance band cuts for the interim assessments have either been set too high or too low (depending on content area and grade) relative to the CSAP **Use 3** • Different rates of effectiveness found for many classrooms when comparing growth rates achieved by students on interim tests and CSAP • The raw scores can understate the amount of growth achieved by students between the version 1 and 3 tests. In this study, the raw scores understated growth for the majority of grade 4 classrooms reviewed on the reading tests

Figure 38. Evaluating student growth for teacher merit pay: summary of strengths and weaknesses.

As indicated in Chapter 5, the extent to which growth rates between a state assessment and other assessments used to evaluate student growth should be correlated for teacher merit pay purposes is a topic of current debate. This study's finding of a moderate correlation (.47) for math and a low correlation (.33) for reading indicates that many teachers would receive different effectiveness ratings depending on which test was being used to evaluate student growth. Considering that a state summative test is furnished with at least twice as many items than most interim assessments, the growth rates evaluated using the state test would likely evaluate student performance better than many interim tests. The variability in effectiveness ratings found in this study is especially disconcerting as more states intend to use interim tests to evaluate teachers. Further, since most teachers instruct in subjects where there is no state test available to compare student growth attributed to teacher effectiveness using these interim tests, these findings call for ensuring that additional measures are included to evaluate the performance of teachers in non-state tested subjects.

Use 4	Strengths	Criticisms
Improve instruction to meet the learning needs of students	**Inference 1** • Most tests are reliable at .8 and above	**Inference 1** • Low reliabilities (alpha < .5) for standards populated with fewer than five items **Inference 2** • Interviews suggest that general (standards level) rather than diagnostic information can be inferred about students. Further, interviewed teachers only make instructional decisions if all students miss a specific standard not taught or adequately covered in previous lessons • Responses from the survey indicate that 44% of teachers agree that they can identify student misunderstandings and errors on the interim assessments **Inference 3** • Interviews suggest that the data are not used to help inform their contextual understanding of individual students, but are used largely to understand whole classroom performance on broader standards • Responses from the survey indicate that 41% of teachers agree that the interim assessments give a good indicator of what students are learning in the classroom **Use 4** • Interviews suggest that teachers do not use the assessments to provide targeted instruction for students • Responses from the survey indicate that 46% of teachers agree that the interim assessments serve as a useful tool for helping students identify what they know and what they still need to learn

Figure 39. Evaluating use of assessments to improve instruction: summary of strengths and weaknesses.

Although the interim tests reflected upon by the interviewed teachers and the population of teachers surveyed are based on the tests produced by TPR, and the content depth of these interim tests may have improved in subsequent years, two primary issues should still be considered by this district. The first issue concerns the reporting of subscores with fewer than five items reported. Considering that subscores with fewer than five items reported are likely not providing reliable information about students, these subscores should not be reported out as separate areas for teachers to make “diagnostic assessments” about student performance. The second issue concerns the lack of professional development opportunities to help teachers integrate the use of interim assessments to improve student learning. Assuming that the current interim assessments consist of items that reflect more content depth and are better aligned with the curriculum than the assessments evaluated in this study, ProComp staff recently indicated that training is still lacking to help teachers both interpret and use the data to provide targeted instruction to students. As noted by Heritage et al. (2009), the expected use of assessments to improve instruction will not take place if there are inadequate professional development opportunities provided for teachers to learn how to cultivate those practices.

Considering that the way DPS has used interim assessments is not unique to the experience of this school district, unsubstantiated uses of interim assessments are likely to be found in other school districts across the country using interim test programs to meet multiple purposes. In this study, the criticisms noted in each figure can be largely addressed without having to make substantial changes to the assessments. However, the use of these interim assessments to evaluate and reward teachers remains problematic and the findings in this study suggest that the larger educational field needs to consider whether interim assessments can serve as fair and valid measures for evaluating and rewarding teachers.

Since many states, including Colorado, Delaware, Washington D.C., Rhode Island and North Carolina use, or are considering the use of interim assessments to evaluate teacher performance in both state tested and non-state tested subjects, the findings from this study can provide some useful policy insights. As an example, for non-state tested subjects against which there are no formal measures to compare interim measures, the district or state should seriously question whether growth assessed by interim tests should be given as much weight in a teacher's performance evaluation relative to other measures. As evidenced by this study, if interim measures designed to capture the same power standards as a state test can exhibit different teacher effectiveness ratings than the state test, decision makers should not automatically assume that these instruments could adequately evaluate effectiveness for non-state tested subjects. Unless adequate evidence can be presented that the selected interim test provides a good measure of teacher effectiveness in non-state tested as well as state tested subjects, the findings from this study should caution against placing considerable weight in a teacher's evaluation using growth measures from these assessments. Re-evaluating the weighting scheme of a teacher's performance evaluation should also be considered especially by states such as Colorado, that plan to attribute fifty percent of a teacher's performance evaluation on growth measures. For teachers without a state-test to evaluate their performance, some districts and states may be tempted out of convenience to use interim assessments as common growth measures of evaluating teacher performance across a given content area.

Limitations and Future Research

As addressed in Chapter 3, the appraisal of the interpretive arguments was limited in scope and left certain inferences and assumptions unchecked. Evaluating inferences such as Content Design and Extrapolation would require entirely different types of studies and data to determine the extent to which those inferences and assumptions hold. For example, under Content Design, curriculum experts can evaluate the extent to which the content of the interim test administered at the end of each testing period is aligned with a specific grade and content's scope and sequence. This content evaluation could be conducted to test the assumption that the interim assessment adequately represents the depth of content taught before the test administration date. This type of study would be especially relevant to any district considering using interim tests for pay for performance purposes since Lockwood et al. (2007) caution against evaluating teacher effectiveness using tests that are not reflective of the scope and sequence represented in the subject taught by the teacher. If the assumption checks of the Content Design Inference reveal that the test does not appear to adequately reflect or align with the scope and sequence of the content taught prior to the testing period, then this finding would mean that all other areas of the interpretive argument evaluated earlier are no longer valid. In other words, if the content that was supposed to be assessed by the interim test is not adequately represented, then inferences can no longer be drawn about using the test to evaluate student mastery or knowledge over the target domain assessed. For this district and for all other districts at the national level, a more in-depth content review would be useful as a first step in establishing whether the interim assessment serves as an appropriate instrument for evaluating teacher effectiveness.

Another limitation in this study was that the assumptions underlying the first two “Use” inferences (see assumptions under Figure 34 for Use 1 and Use 2) were not checked. Evaluating how well the uses are benefitting those directly impacted by the policy decision serves as an important validity check. For Use 1, the assumption was made by the former CAO that students identified for mandatory remediation in the summer of 2007 would receive the academic skills needed to better succeed in the high school environment. Although a study of whether the services received appeared to better prepare these students for high school goes beyond an examination of the interim assessments, this type of study is important to ensure that these students actually benefitted from this use.

Another area for future research that falls outside the scope of this dissertation study is to conduct a cost-benefit analysis of an interim assessment program. As noted by Olsen (2005), the cost of interim assessment programs is considerable for any school district or state. In the experience of this school district and all other school districts nation-wide, costs go well beyond the initial purchase of items or the product from a vendor. In addition, since all districts have to provide all school sites with training in the administration, scoring, and interpretation of the test data, staff costs should be factored into an overall financial analysis of an interim assessment program. These costs would then need to be assessed against the benefits of the program.

The benefit or ultimate goal of any assessment program is to improve student learning and student achievement (Pellegrino, et al., 2001). Many school districts across the nation invest in interim assessment systems, firmly believing that these systems aid in improved achievement and learning (Shepard, 2007; Herman & Baker, 2005; Olsen, 2005). Currently, the extent to which these systems are improving achievement remains questionable at best. The last two studies (see Nunnery et al., 2003 and Henderson, et al., 2007, 2008) reviewed in Chapter 3 present mixed findings of student achievement improvements detected after interim assessments were introduced to each site. For all districts, an assessment of costs relative to the extent to which student achievement appears to be improving from the interim assessment data would provide useful evidence for all district stakeholders. Considering that almost all districts nation-wide currently face large budget shortfalls, the cost and benefit of interim test programs should also be evaluated against other line items considered for cuts in many school districts and state departments of education (e.g., teacher layoffs, reducing the paraprofessional workforce, and eliminating after school programs or other types of student services).

Conclusion

Many districts across the U.S. have invested in interim test programs to fulfill evaluative, predictive and instructional purposes. Many more districts plan to do so based on the perceived effectiveness of these tests to fulfill those purposes. The findings from this dissertation study reveal that although the predictive claim of interim assessments appears to be well founded the use of interim assessments as measures of teacher effectiveness measures and as tools for instructional improvement may be problematic. The validity evidence gathered to support the use of interim tests to identify students for mandatory remediation is also questionable, but unlike the merit pay and instructional uses, staff can improve upon the precision of the test and re-evaluate the number of cuts set on the test without having to make any substantive changes to the assessment.

While there is a clear trend in school districts across the country to use interim assessments, the key findings from this study reveal that these assessments may not always be providing accurate information to drive aspects of accountability and reforms such as evaluating teacher effectiveness, school performance, and improving instruction. The education sector's decision to use interim assessments as accountability and reform tools means that these assessments can have a long-term effect on the educational system. That is, in addition to costs diverted to interim assessments, the data from these assessments could, for better or for worse, be used over time as evidence used to identify effective or ineffective teachers or as evidence of how well students have mastered important skills and knowledge. If the goals of interim assessments are deemed important by district and state department of education leaders, significant time and energy needs to be dedicated to establish what these tests can be used for and to evaluate how well these tests are providing stakeholders with valid information required to support both accountability decisions and reforms. Without providing guidance on appropriate uses and evaluating the extent to which the selected interim assessment program provides valid information to drive important decisions both at the classroom and the larger systems level, educational practitioners need to consider the prospect that these assessments may do more long-term harm than good to both students and teachers.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. New York: AERA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees' ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores*. pp. 397-479. Reading, MA: Addison-Wesley.

Black, P. and William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5(1), 7-74.

Blanc, S., Christman, J.B., Hugh, R., Mitchell, C., & Travers, E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, 85(2), 205-225.

Brennan, R. (1992). An NCME Instructional Module on Generalizability Theory. *Instructional Topics in Educational Measurement, 225-232*.

Brennan, R. L. (1995). *The Conventional Wisdom About Group Mean Scores. Journal of Educational Measurement*, 32, 385-396.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical scale. *Education Finance & Policy*, 4(4), 384-414.

Briggs, D., & Wilson, M. (2003). An Introduction to Multidimensional Measurement Using Rasch Models. *Journal of Applied* (Yeh, 2006)*Measurement*, 4(1), 87-100.

Buckley, K., & Marion, S. (2011). *A Survey of Approaches Used to Evaluate Educators in Non-Tested Grades and Subjects*. Unpublished manuscript.

Bulkley, K., Christman, J., Goertz, M., & Lawrence, N. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education*, 85(2), 186-204.

Cech, S.J. (2008). Test industry split over "formative" assessment. *Education Week*, 28(4), 1-15.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29, 3-13.

Clune, W.H., & White, P.A. (2008). *Policy Effectiveness of Interim Assessments in Providence Public Schools*. WCER Working Paper No. 2008-10, Wisconsin Center for Education Research. Madison: University of Wisconsin.

Craig, C.J. (2004). The Dragon in School Backyards: The Influence of Mandated Testing on School Contexts and Educators' Narrative Knowing, *Teachers College Record*, Volume 106, Number 6, pp, *1229-1257*.

Cronbach, L. J. (1971). Test Validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Darling-Hammond, L. (1989). Standards, Accountability, and School Reform, *Teachers College Record* Volume 106, Number 6, pp. 1047-1085.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum, Associates, Inc.

Halverson, R. (2010). School Formative Feedback Systems. *Peabody Journal of Education*, (85), 130-146.

Hanson, B. A. (2002). IRT Command Language. Available at: www.bah.com/software/irt/icl/index.html.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *"Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2007-No. 039 ed.). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2008-No. 002 ed.). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24-31.

Herman, J. L., & Baker, E. L. (2005). Making Benchmark Testing Work. *Educational Leadership*, 63(3), 48-54.

Herman, J. L., Yamashiro, K., & Lefkowitz, S. (2008). *Exploring Data Use and School Performance in an Urban Public School District Evaluation of Seattle Public Schools' Comprehensive Value-Added Assessment System.* CRESST Report 742.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.

Kane, M.T. (1992). An arguments-based approach to validation. *Psychological Bulletin*, (112), 527-535.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-63). New York: American Council on Education.

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking Methods and Practices* (2nd Edition ed.). New York: Springer.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Le, V., & Martinez, F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement*, *44(1):47-67*.

Masters, G. N. (1992). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-74.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-76.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy of Sciences.

Division of Assessment and Accountability, New Mexico Public Education Department. (2006). *Consumer Guide to Formative Assessments.* Retrieved from: http://www.ped.state.nm.us/div/acc.assess/assess/dl/Formative_assessment_consumer_guide/Consumer%20Guide%20Final.pdf

Neild, R.C. (2008). Instructional Guidance, Collegial Work: Evidence on Teachers' Use of Benchmarks from a Large-Scale Survey. Paper presented at the annual meeting of the American Educational Research Association (March, 2008): New York, New York.

Niemi, D., Vallone, J., Wang, J., & Griffin, N. (2007, July). *Recommendations for building a valid benchmark assessment system: Interim report to the Jackson Public Schools* (CRESST Report 723 ed.). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Office of Assessment, Research and Data Analysis, Miami-Dade County Public Schools. (2008). *How Interim Assessments Affect Student Achievement*. Information Capsule, Vol. 0804. Retrieved from: http://drs.dadeschools.net/InformationCapsules/IC0804.pdf

Oláh, L.N., Lawrence, N.R., & Riggan, M. (2010). *Learning to learn from benchmark assessment data: How teachers analyze results*. Peabody Journal of Education 85: 226-245.

Olson, L. (2005a). Benchmark Assessments Offer Regular Checkups on Student Achievement. *Education Week*, *25*(13), 13-14.

Olson, L. (2005b). Not All Teachers Keen on Periodic Tests. *Education Week*, *25*(13), 13.

Perie, M., Marion, S., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5-13.

Nunnery, J., Ross, S. M., & Goldfeder, E. (2003). *The effect of School Renaissance on TAAS scores in the McKinney ISD*. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Rothstein, J (2011). Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project." Boulder, CO: National Education Policy Center. Retrieved February 10, 2011 from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*: *Vol. 1. Measurement methods for the social sciences series*. Newbury Park, Calif.: Sage Publications.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed), *Review of Research in Education* (Vol. 19, pp. 405-450). Washington DC: American Educational Research

Shepard, L. A. (2007). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279-303). Mahwah, NJ: Lawrence Erlbaum Associates.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, 28(3), 32-37.

Thissen, D., & Wainer, H. (2001). *(Eds)*, Test Scoring. Hillsdale, NJ: Lawrence Erlbaum Associates.

U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* (GAO/PEMD-93-12 ed.). Washington, DC: Author.

Vendlinski, T.P., Nagashima, S., & Herman, J.L. (2007). *Creating accurate science benchmark assessments to inform instruction*. Los Angeles, CA: CRESST. Retrieved from http://www.cse.ucla.edu/products/reports/R730.pdf.

Villano, M. (2006). "Assessing Formative Assessment.". In *Technology & Learning*. Dayton: Vol. 26, (6); pg. 8-11.

Vygotsky, L.S. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Walker, C. M., & Beretvas, S.N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, 40, 255-275.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability Theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science, pp. 717-719*. New York: Wiley.

Weeks, J.P. (2007). Plink: IRT separate calibration linking methods (R package version 0.0-4). Available. Retrieved from http://cran.r-project.org/web/packages/plink/index.htm.

Wilson, M. (2005). *Constructing Measures, An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum, Associates, Inc.

Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching. *Washington: Bill & Melinda Gates Foundation, 1*. Retrieved December 16, 2010, from www. metproject.org/downloads/met-framing-paper.pdf.

Wu, M., & Adams, R.J. (In Press). Properties of rasch residual fit statistics. *Journal of Applied Measurement*.

Yeh, S. S. (2006). Can rapid assessment moderate the consequences of high-stakes testing? *Education and Urban Society*, 39(1), 91-112.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 20, 71-88.

Appendix A

Appendix A-1. Example of an individual scoring report

Overall Performance:

Percent of students at
Proficient
or above:

15%

Average Score: 11.7/30 (39%)

Per Band Performance:

Band	Range	# Students	%
Unsatisfactory	0.00-9.00	11	42%
Partially Proficient	9.01-16.00	11	42%
Proficient	16.01-24.00	4	15%
Advanced	24.01-30.0	0	0%

Per Standard Performance:

Standard	Below Profiency	At or Above Profiency
COAF--Mathematics (2004): Standard 1	18 (69%)	8 (31%)
COAF--Mathematics (2004): Standard 2	24 (92%)	2 (8%)
COAF--Mathematics (2004): Standard 3	6 (23%)	20 (77%)
COAF--Mathematics (2004): Standard 4	24 (92%)	2 (8%)
COAF--Mathematics (2004): Standard 5	16 (62%)	10 (38%)
COAF--Mathematics (2004): Standard 6	25 (96%)	1 (4%)
COAF: Other	26 (100%)	

Question Group Performance:

Question Group	Below Proficiency	At or Above Proficiency
Framework Statement 1.1a	19 (73%)	7 (27%)
Framework Statement 1.1b	8 (31%)	18 (69%)
Framework Statement 1.2a	10 (38%)	16 (62%)
Framework Statement 1.2b	17 (65%)	9 (35%)
Framework Statement 1.4a	21 (81%)	5 (19%)
Framework Statement 2.1a	16 (62%)	10 (38%)
Framework Statement 2.2a	21 (81%)	5 (19%)
Framework Statement 2.3a	8 (31%)	18 (69%)
Framework Statement 2.5a	18 (69%)	8 (31%)
Framework Statement 2.5b	23 (88%)	3 (12%)
Framework Statement 3.1b	7 (27%)	19 (73%)
Framework Statement 3.2a	9 (35%)	17 (65%)
Framework Statement 3.4a	6 (23%)	20 (77%)
Framework Statement 4.3a	20 (77%)	6 (23%)
Framework Statement 4.5b	13 (50%)	13 (50%)
Framework Statement 4.5c	24 (92%)	2 (8%)
Framework Statement 5.2a	17 (65%)	9 (35%)
Framework Statement 5.3b	2 (8%)	24 (92%)
Framework Statement 5.5a	22 (85%)	4 (15%)
Framework Statement 6.2a	22 (85%)	4 (15%)
Framework Statement 6.2b	22 (85%)	4 (15%)
Framework Statement 6.3b	18 (69%)	8 (31%)
Framework Statement 6.4b	26 (100%)	

Appendix B

Appendix B-1. Benchmarks for Grade 8 by Standard

Standard 1 Benchmarks

1. Demonstrate Meanings for integers, rational numbers, percents, exponents, square roots, and pi, use physical materials and technology in problem solving situations;
2. Read, write and order integers, rational numbers, and common irrational numbers;
3. Apply number theory concepts (for example, primes, factors, multiples) to represent numbers in various ways;
4. Use the relationships among fractions, decimals and percents include the concepts of ratio and proportion, in problem-solving situations;
5. Develop, test, and explain conjectures about properties of integers and rational numbers; and
6. Use number sense to estimate and justify the reasonableness of solutions to problems involving integers, rational numbers, and common irrational numbers.

Standard 2 Benchmarks

1. Represent, describe, and analyze patterns and relationships using tables, graphs, verbal rules, and standard algebraic notation;
2. Describe patterns using variables, expressions, equations, and inequalities in problem-solving situations;
3. Analyze functional relationships to explain how a change in one quantity results in a change in another (for example, how the area of a circle changes as the radius increases, or how a person's height changes over time);
4. Distinguish between linear and nonlinear functions through formal investigations; and
5. Solve simple linear equations in problem-solving situations using a variety of methods (informal, formal, graphical) and a variety of tools (physical materials, calculators, computers).

Standard 3 Benchmarks

1. Read and construct displays of data using appropriate techniques (for example, line graphs, circle graphs, scatter plots, box plots, stem-and-leaf plots) and appropriate technology;
2. Display and use measures of central tendency, such as mean, median, and mode, and measures of variability such as range and quartiles;
3. Evaluate arguments that are based on statistical claims;
4. Formulate hypotheses, draw conclusions, and make convincing arguments based on data analysis;
5. Determine probabilities through experiments or simulations;
6. Make predictions and compare results using both experimental and theoretical probability drawn from real world problems; and
7. Use counting strategies to determine all the possible outcomes from an experiment (for example, the number of ways students an line up to have their picture taken).

Standard 4 Benchmarks

1. Construct two-and three-dimensional models using a variety of materials and tools;
2. Describe, analyze, and reason informally about the properties (for example, parallelism, perpendicularity, congruence) of two- and three- dimensional figures;

3. Apply the concepts of ratio, proportion, and similarity in problem-solving situations;
4. Solve problems using coordinate geometry;
5. Solve problems involving perimeter and area in two dimensions and involving surface area and volume in three dimensions; and
6. Transform geometric figures using reflections, translations, and rotations to explore congruence.

Standard 5 Benchmarks

1. Estimate, use, and describe measures of distance, perimeter, area, volume, capacity, weight, mass, and angle comparison;
2. Estimate, make and use direct and indirect measurements to describe and make comparisons;
3. Read and interpret various scales including those based on number lines, graphs, and maps;
4. Develop and use formulas and procedures to solve problems involving measurement;
5. Describe how a change in an object's linear dimensions affects its perimeter, area, and volume; and
6. Select and use appropriate units and tools to measure to the degree of accuracy required in a particular problem-solving situation.

Standard 6 Benchmarks

1. Use models to explain how ratios, proportions and percents can be used to solve real-world problems;
2. Construct, use and explain procedures to compute and estimate with whole numbers, fractions, decimals, and integers;
3. Develop, apply, and explain a variety of different estimation strategies in problem-solving situations and explain why an estimate may be acceptable in place of an exact answer; and
4. Select and use appropriate algorithms for computing with commonly used fractions and decimals, percents, and integers in problem-solving and determine whether the results are reasonable.

Appendix C

Appendix C-1. Point-Biserials for Selected Interim Assessments (Sorted by Grade, Content Area and Administration Date)

Grade 4, Version 1 Math 2007

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.52	0.52			
2	-0.44	0.44			
3	-0.32	0.32			
4	-0.41	0.41			
5	-0.42	0.42			
6	-0.55	0.55			
7	-0.4	0.4			
8	-0.44	-0.1	0.51		
9	-0.45	0.45			
10	-0.5	0.5			
11	-0.4	0.4			
12	-0.45	0.45			
13	-0.34	0.34			
14	-0.36	0.36			
15	-0.44	0.44			
16	-0.38	-0.23	0.19	0.53	
17	-0.24	0.24			
18	-0.5	0.5			
19	-0.48	0.48			
20	-0.21	0.21			
21	-0.44	0.44			
22	-0.52	0.52			
23	-0.57	-0.02	0.19	0.31	0.44

Grade 4, Version 3 Math 2008

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.44	0.44			
2	-0.47	0.47			
3	-0.49	0.49			
4	-0.67	-0.07	0.15	0.25	0.57
5	-0.3	0.3			
6	-0.64	-0.09	0.67		
7	-0.47	0.47			
8	-0.54	0.54			
9	-0.45	0.45			
10	-0.36	0.36			
11	-0.47	0.47			
12	-0.46	-0.29	0.17	0.53	
13	-0.51	0.51			
14	-0.38	0.38			
15	-0.46	0.46			
16	-0.57	0.57			
17	-0.49	0.49			
18	-0.62	0.17	0.57		
19	-0.37	0.37			
20	-0.38	0.38			
21	-0.4	0.4			
22	-0.44	0.44			
23	-0.54	0.54			

Grade 4, Version 1 Reading 2007

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.45	0.45			
2	-0.58	0.58			
3	-0.52	0.52			
4	-0.44	0.44			
5	-0.51	0.51			
6	-0.55	0.55			
7	-0.43	0.43			
8	-0.45	-0.28	0.12	0.5	
9	-0.57	0.32	0.38		
10	-0.42	0.42			
11	-0.41	0.41			
12	-0.45	0.45			
13	-0.59	0.59			
14	-0.4	0.4			
15	-0.5	0.5			
16	-0.63	0.28	0.48		
17	-0.44	0.44			
18	-0.23	0.23			
19	-0.34	0.34			
20	-0.39	0.39			
21	-0.43	0.43			
22	-0.43	0.43			
23	-0.39	0.39			
24	-0.58	0.19	0.35	0.33	
25	-0.41	0.3	0.28		

Grade 4, Version 3 Reading 2008

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.56	0.56			
2	-0.46	0.46			
3	-0.45	0.45			
4	-0.41	0.41			
5	-0.47	0.47			
6	-0.53	0.11	0.42		
7	-0.54	0.54			
8	-0.51	0.51			
9	-0.5	-0.27	0.13	0.47	
10	-0.51	0.51			
11	-0.46	0.46			
12	-0.46	0.46			
13	-0.46	0.46			
14	-0.45	0.45			
15	-0.56	0.56			
16	-0.24	0.24			
17	-0.5	-0.08	0.46		
18	-0.58	-0.24	0.03	0.57	
19	-0.53	0.53			
20	-0.29	0.29			
21	-0.55	0.01	0.51		
22	-0.5	0.5			
23	-0.57	0.57			
24	-0.37	0.37			
25	-0.49	0.49			

Grade 8, Version 2 Math 2007

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.3	0.3			
2	-0.47	0.47			
3	-0.37	0.37			
4	-0.48	0.48			
5	-0.45	0.45			
6	-0.24	0.12	0.26		
7	-0.37	0.37			
8	-0.44	0.44			
9	-0.4	0.4			
10	-0.47	0.47			
11	-0.45	0.45			
12	-0.52	0	0.43	0.31	
13	-0.46	0.46			
14	-0.49	0.49			
15	-0.18	0.18			
16	-0.48	0.48			
17	-0.47	0.47			
18	-0.55	-0.08	0.19	0.39	0.34

Grade 8, Version 2 Writing 2007

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.5	0.5			
2	-0.48	0.48			
3	-0.51	0.51			
4	-0.55	0.55			
5	-0.45	0.45			
6	-0.56	0.56			
7	-0.51	0.51			
8	-0.42	0.42			
9	-0.47	0.47			
10	-0.34	0.34			
11	-0.45	0.45			
12	-0.41	0.41			
13	-0.6	0.6			
14	-0.51	0.51			
15	-0.52	0.52			
16	-0.53	0.53			
17	-0.51	-0.33	-0.08	0.37	0.39
18	-0.53	-0.35	0.03	0.42	0.35
19	-0.55	-0.1	0.51		

Grade 8, Version 2 Writing 2008

	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.37	0.37			
2	-0.11	0.11			
3	-0.05	0.05			
4	-0.35	0.35			
5	-0.42	0.42			
6	-0.35	0.35			
7	-0.38	0.38			
8	-0.39	0.39			
9	-0.14	0.14			
10	-0.41	0.41			
11	-0.31	0.31			
12	-0.51	0.51			
13	-0.58	0.58			
14	-0.47	0.47			
15	-0.51	0.51			
16	-0.45	0.45			
17	-0.59	-0.24	0.03	0.42	0.39
18	-0.57	-0.31	0.07	0.47	0.36
19	-0.6	-0.04	0.55		

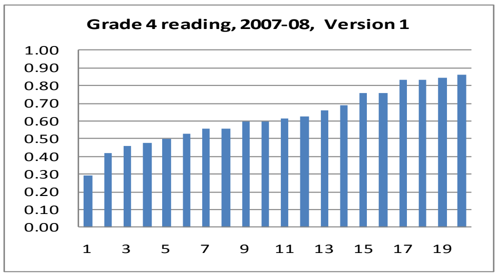
Grade 10, Version 2 Math 2007

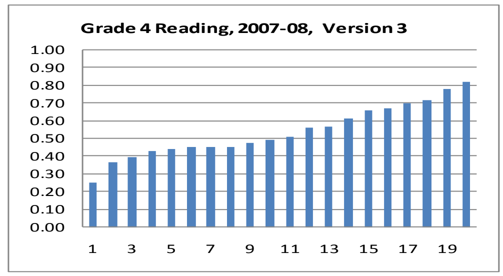
	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.51	0.51			
2	-0.56	0.07	0.22	0.42	0.37
3	-0.52	0.52			
4	-0.5	0.5			
5	-0.49	0.49			
6	-0.61	0.06	0.28	0.55	
7	-0.21	0.21			
8	-0.45	0.45			
9	-0.41	0.41			
10	-0.45	0.45			
11	-0.47	0.47			
12	-0.25	0.25			
13	-0.44	0.44			
14	-0.4	0.4			
15	-0.55	0.55			
16	-0.41	0.41			
17	-0.49	0.49			

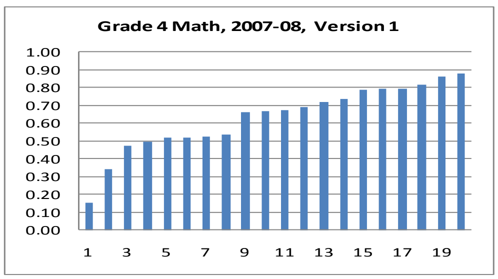
Grade 10, Version 2 Reading 2007

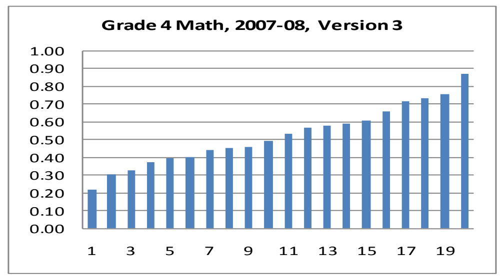
	RESPONSE CATEGORIES				
Item #	0	1	2	3	4
1	-0.49	0.49			
2	-0.62	-0.1	0.25	0.48	
3	-0.39	0.39			
4	-0.53	0.53			
5	-0.37	0.37			
6	-0.4	0.4			
7	-0.56	-0.08	0.55		
8	-0.57	0.57			
9	-0.31	0.31			
10	-0.34	0.34			
11	-0.27	0.27			
12	-0.61	-0.07	0.32	0.47	
13	-0.27	0.27			
14	-0.5	0.5			
15	-0.41	0.41			
16	-0.48	0.48			
17	-0.55	0.22	0.45		
18	-0.48	0.16	0.38		
19	-0.1	0.1			
20	-0.4	0.4			
21	-0.5	0.5			
22	-0.43	0.43			
23	-0.33	0.33			
24	-0.26	0.26			
25	-0.38	0.38			
26	-0.48	0.48			

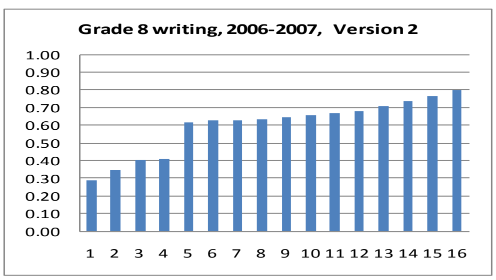
Appendix C-2. Distribution of p-values for nine interim tests

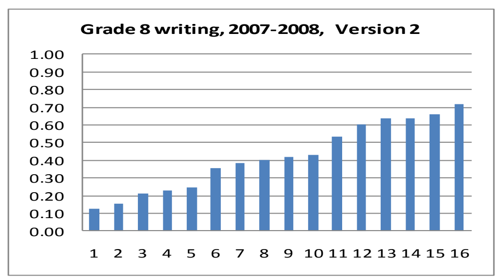


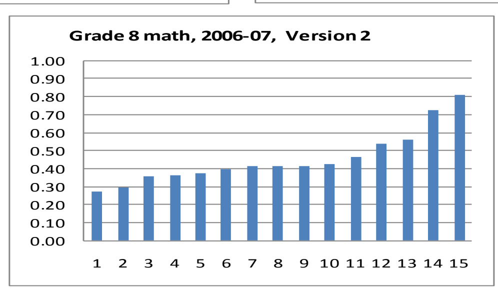


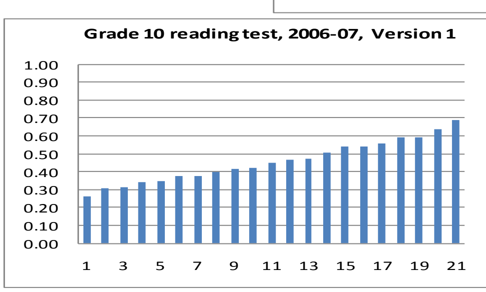


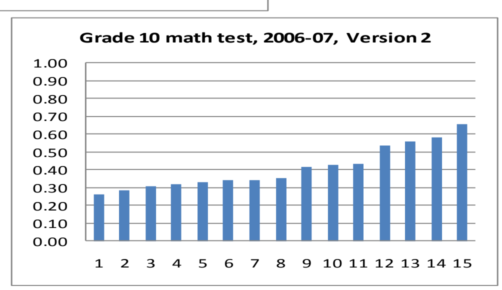




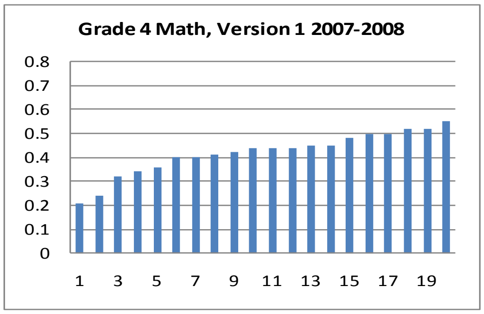


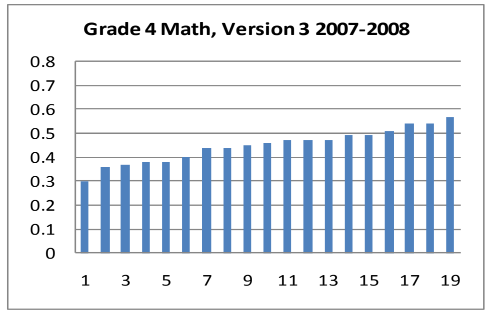


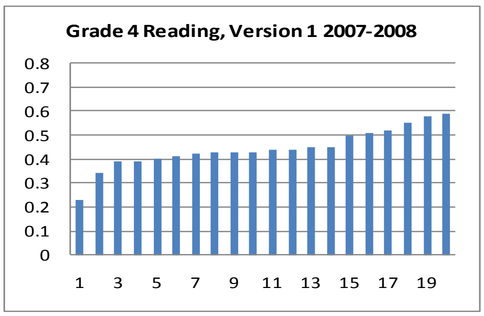


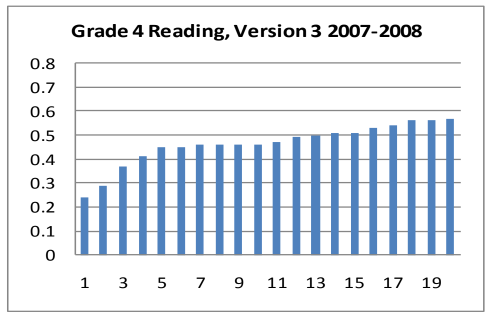


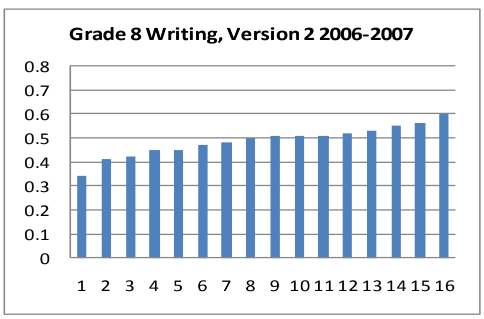
Appendix C-3. Distribution of MC items point-biserials for nine interim tests

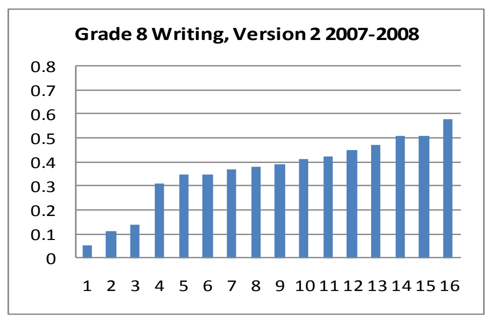


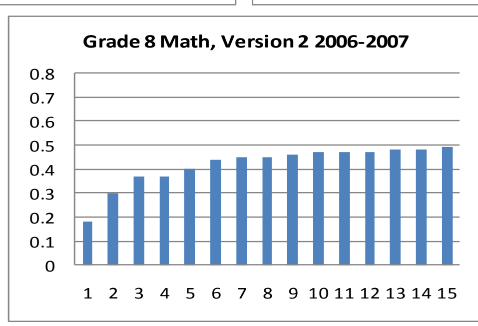


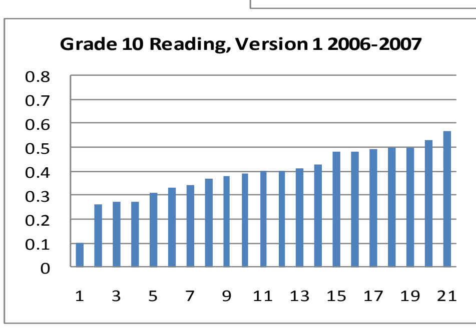


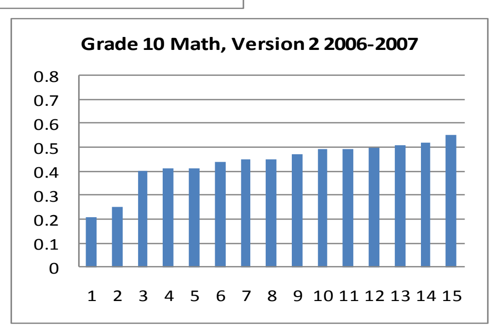












Appendix C-4. An example of an interim assessment scoring report for a student

Student Performance Report

Date: February 12, 2009
District Student ID:
School:
Grade: 8

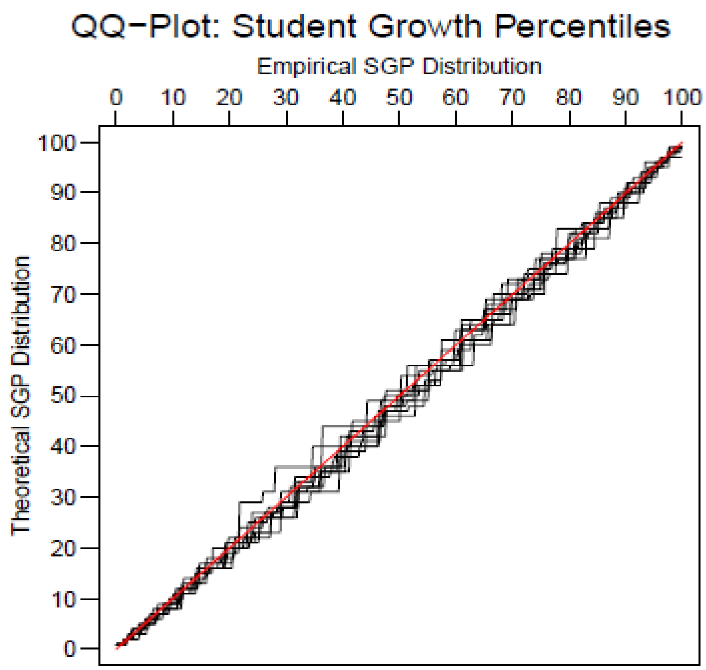
Benchmark Assessments	Admin Date	Last Score Date	Performance Band	Raw Score	% Correct
7th Gr Math Test 3	April 2008	05/13/2008	Partially Proficient	12	40 %
COAF--Mathematics (2004): Standard 1			Below Profiency	1	17 %
COAF--Mathematics (2004): Standard 2			At or Above Profiency	6	75 %
COAF--Mathematics (2004): Standard 3			At or Above Profiency	3	100 %
COAF--Mathematics (2004): Standard 4			Below Profiency	0	0 %
COAF--Mathematics (2004): Standard 5			At or Above Profiency	2	67 %
COAF--Mathematics (2004): Standard 6			Below Profiency	0	0 %
COAF: Other			Below Profiency	0	0 %
Framework Statement 1.1a			Below Proficiency	0	0 %
Framework Statement 1.1b			Below Proficiency	0	0 %
Framework Statement 1.2a			Below Proficiency	0	0 %
Framework Statement 1.2b			At or Above Proficiency	1	100 %
Framework Statement 1.4a			Below Proficiency	0	0 %
Framework Statement 2.1a			At or Above Proficiency	3	75 %
Framework Statement 2.2a			Below Proficiency	0	0 %
Framework Statement 2.3a			At or Above Proficiency	1	100 %
Framework Statement 2.5a			At or Above Proficiency	1	100 %
Framework Statement 2.5b			At or Above Proficiency	1	100 %
Framework Statement 3.1b			At or Above Proficiency	1	100 %
Framework Statement 3.2a			At or Above Proficiency	1	100 %
Framework Statement 3.4a			At or Above Proficiency	1	100 %
Framework Statement 4.3a			Below Proficiency	0	0 %
Framework Statement 4.5b			Below Proficiency	0	0 %
Framework Statement 4.5c			Below Proficiency	0	0 %
Framework Statement 5.2a			At or Above Proficiency	1	100 %
Framework Statement 5.3b			At or Above Proficiency	1	100 %
Framework Statement 5.5a			Below Proficiency	0	0 %

Appendix D

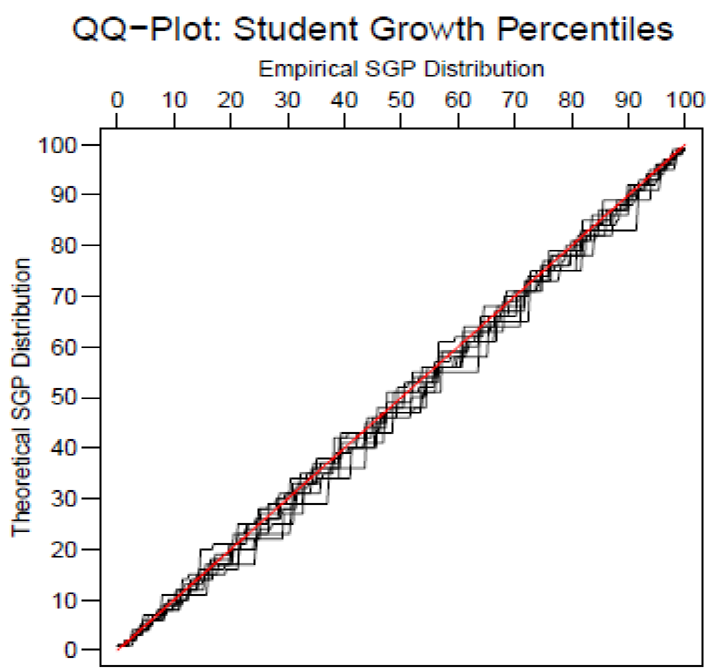
Appendix D-1. ProComp Merit Pay and Salary Increase Schedule by Component

Component of Index $35,568	Knowledge and Skills			Comprehensive Professional Evaluation		Market Incentives		Student Growth		
Element	Professional Development	Advanced Degree and License	Tuition Reimbursement	Probationary	Non-Probationary	Hard to Serve School	Hard to Staff Assignment	Student Growth Objectives	Exceeds CSAP Expectations	Distinguished Schools
Description of Element	Providing ongoing professional development – tied to the needs of our students -- is a central strategy to help you expand your skills, improve student performance, and advance your career with the district	Compensation for Graduate Degree or Advanced Licenses or Certificates	Reimbursement for tuition.	Increases for new teachers based on a satisfactory evaluation.	Increases based on a satisfactory evaluation.	Designed to attract teachers to schools with a high free and reduced lunch percentage.	Designed to attract teachers to roles with high vacancy rate and high turnover	Incentive paid for meeting student growh objectives.	Teachers whose assigned student's growth in CSAP scores exceed district expectations	Incentive is paid to teachers in schools recognized for outstanding performance
Eligibility and Payout	Base building for PDU earned. Additional PDUs may be banked with no limit (only pay one per year)	Paid upon receipt of documentation that the license or certification is active and current	Paid upon receipt of evidence of payment for and satisfactory completion of coursework; $1000 lifetime account	Requires Satisfactory Evaluation: If unsatisfactory, ineligible for CPE increase	Requires Satisfactory Evaluation: If unsatisfactory, ineligible for CPE increase.	Teachers currently serving in schools designated "Hard-to-Serve".	Teachers currently serving in designated "Hard-to-Staff" positions	Base building when 2 SGOs are met, non base-building when only 1SGO is met during prior school year	Sustainable increase paid for exceeding expectations; sustainable decrease for falling below expectations	Paid based on performance during the prior school year.
Affect on Base Salary	Base Building	Base Building	Non-Base Building	Base Building	Base Building	Non-Base Building	Non-Base Building	Base Building/Non-Base Building	Sustainable Base Building	Non-Base Building
Percent of Index	2%	9% per degree or license. Eligible once every 3 yrs	N/A	1% every year	3% every three years	3.0%	3.0%	1%	3.0%	2.0%
Dollar Amount	$711	$3,201	Actual expense up to $1000 lifetime	$356	$1,067	$1067 $88.92/mo	$1067 ($88.92 per mo) x (# of assignments held)	$356	$1,067	$711
Builds pension and highest average salary	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Payment Type and Frequency	Prorated over 12 months upon submission of poper documentation.	Prorated over 12 months upon submission of poper documentation.	Up to $1000 upon submission of proper documents	Prorated over 12 months. If unsatisfactory delayed at least 1 yr	Prorated over 12 months. If unsatisfactory delayed at least 1 yr	Monthly installment upon completion of service each month	Monthly installment upon completion of service each month	Paid in monthly installments	Prorated over 12 months.	Paid in monthly installments

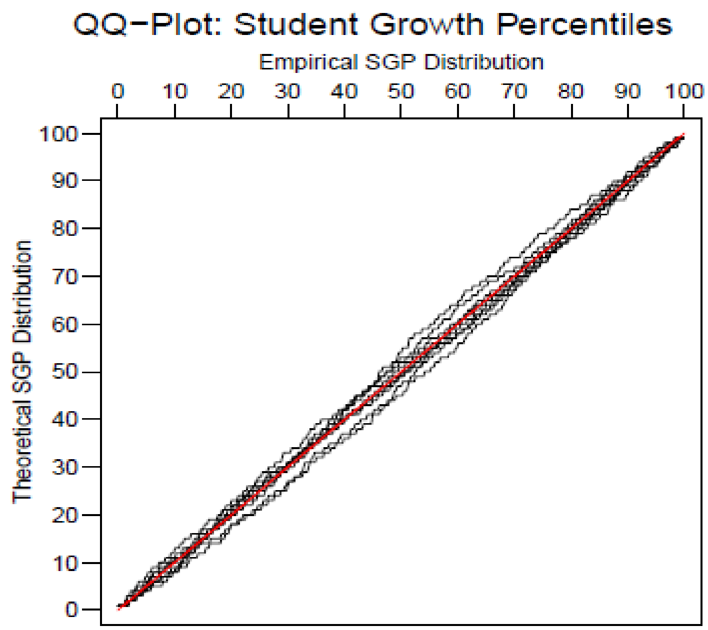
Appendix D-2. Q-Q Plots for Grade 4 Interim and CSAP tests.



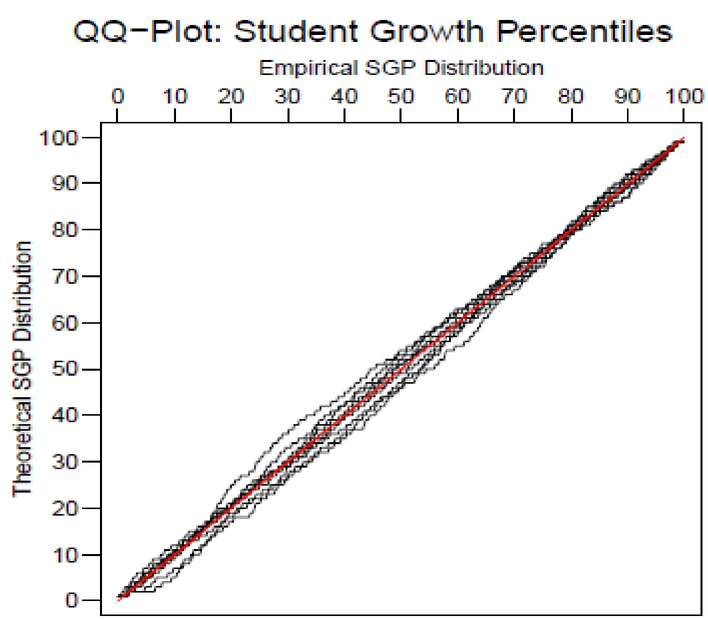
Q-Q Plot for Grade 4 Math Interim Test



Q-Q Plot for Grade 4 Reading Interim Test

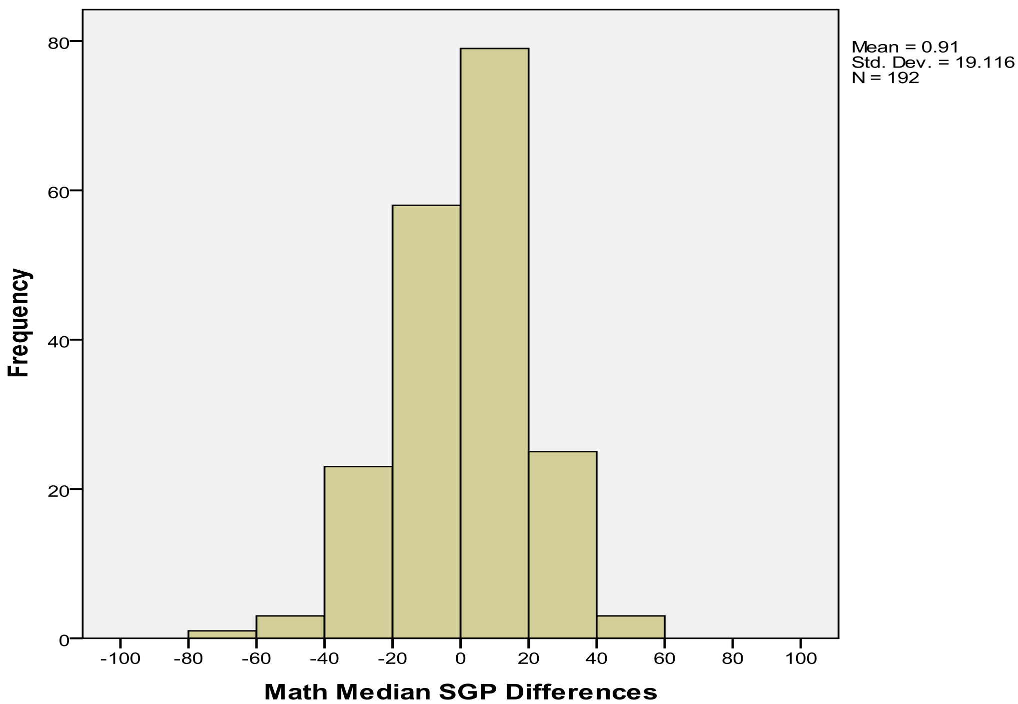


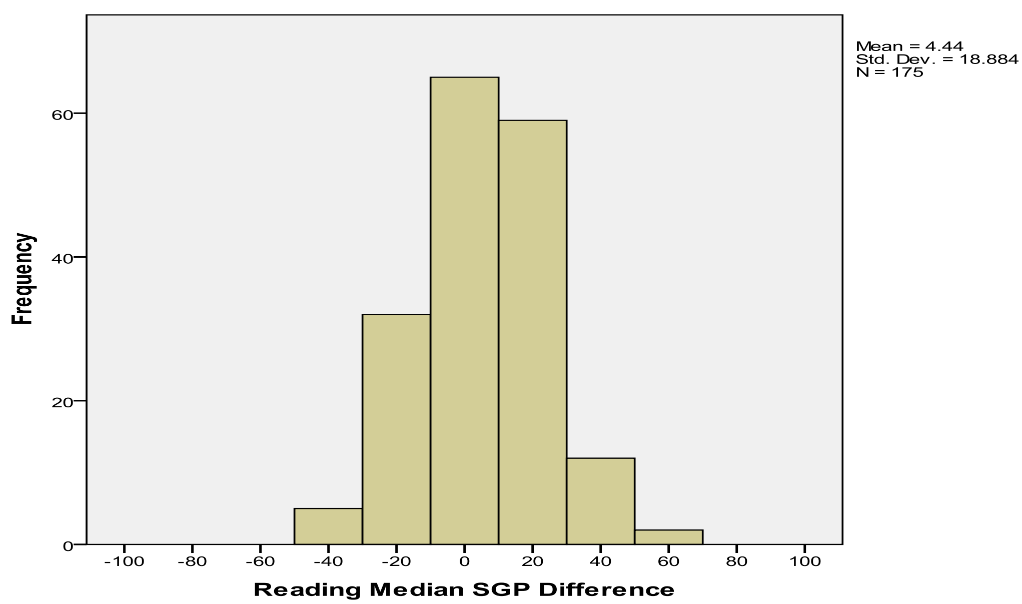
Q-Q Plot for Grade 4 CSAP Math Test



Q-Q Plot for Grade 4 CSAP Reading Test

Appendix D-3. Histograms reflecting Median SGP Differences for Reading and Math Interim Tests and CSAP.





Appendix E

Appendix E-1. Codes developed for reviewing interview transcripts

1. Codes developed before reviewing transcript	2. Codes developed after Review of Transcript
Gaining conceptual insights about student learning (CI)	Evaluating student performance on standards (ST)
Meeting instructional needs of students based on identifying student misconceptions (MI)	Triangulating with other assesments to identify student misconceptions(TRI)
Establishing learning goals with students based on interim assessment data (ELG)	Benefit - gaining insights into testing behavior (Ben_TB)
	Benefit - receiving timely reports compared to CSAP (Ben_Rep)
	Scoring the inteirm tests and length of time used to score (Score)
	Testing overload or too many assessments being used in classroom (Test_Ov)
	Modifying instruction using standards information (Mod_ins)
	Any instructional insights from interim test data (Ins)
	Alignment of content with curriculum and pacing (Align)

Appendix E-2. Survey results by item

Benchmarks set an appropriate pace for teaching to curriculum to my students.

		SD	D	N	A	SA		Not Positive	Neutral	Positive
dem01	Count by Category	270	406	561	562	68	1867	676	561	630
	% by Response	14%	22%	30%	30%	4%	1	36%	30%	34%

(Response categories: SD = Strongly Disagree, D = Disagree, N = Neutral, Agree, SA = Strongly Agree)

Results on the Benchmark tests give me a good indication of what students are learning in my classroom

	SD	D	N	A	SA		Not Positive	Neutral	Positive
Count by Category	254	364	492	687	75	1872	618	492	762
% by Response	14%	19%	26%	37%	4%	100	33%	26%	41%

(Response categories: SD = Strongly Disagree, D = Disagree, N = Neutral, Agree, SA = Strongly Agree)

The Benchmark tests are a useful tool for identifying the content descriptors that students do and do not understand

	SD	D	N	A	SA		Not Positive	Neutral	Positive
Count by Category	214	272	470	863	80	1899	486	470	943
% by Response	11%	14%	25%	45%	4%	100	26%	25%	50%

(Response categories: SD = Strongly Disagree, D = Disagree, N = Neutral, Agree, SA = Strongly Agree)

The Benchmark tests are a useful tool for identifying students' misunderstandings and errors in their reasoning

	SD	D	N	A	SA		Not Positive	Neutral	Positive
Count by Category	229	328	515	768	68	1908	557	515	836
% by Response	12%	17%	27%	40%	4%	100	29%	27%	44%

(Response categories: SD = Strongly Disagree, D = Disagree, N = Neutral, Agree, SA = Strongly Agree)

The Benchmark tests are a useful tool for helping students identify what they know and what they still need to learn.

	SD	D	N	A	SA		Not Positive	Neutral	Positive
Count by Category	234	324	479	793	77	1907	558	479	870
% by Response	12%	17%	25%	42%	4%	100	29%	25%	46%

(Response categories: SD = Strongly Disagree, D = Disagree, N = Neutral, Agree, SA = Strongly Agree)

During the past school year, how often did the following occur in your school?

You examined your students' Benchmarks item analysis.

	1	2	3	4	
Count by Category	168	534	641	354	1697
% by Response	10%	31%	38%	21%	1

(Response categories: 1 = Never, 2 = 1-2 times, 3 = 3-5 times, 4 = More than 5 times, Not Applicable)

Your grade group or coaches met to discuss ideas for re-teaching a skill that students were lacking, according to the Benchmark test.

	1	2	3	4	
Count by Category	402	493	460	319	1674
% by Response	24%	29%	27%	19%	100

(Response categories: 1 = Never, 2 = 1-2 times, 3 = 3-5 times, 4 = More than 5 times, Not Applicable)

Your grade group or coaches met to discuss re-grouping students for instruction on the basis of Benchmarks scores.

	1	2	3	4	
Count by Category	647	404	366	233	1650
% by Response	39%	24%	22%	14%	100

(Response categories: 1 = Never, 2 = 1-2 times, 3 = 3-5 times, 4 = More than 5 times, Not Applicable)